UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author CHOUDHURY

Ratnin

Application and Development of Density Functional Theory

Rathin Choudhury

University College London

A thesis submitted for the degree of

Doctor of Philosophy (Physics)

of the University of London.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration of Originality

I, Rathin Choudhury, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I would like to thank my supervisors Prof. M. J. Gillan and Dr. D. R. Bowler for their support throughout my PhD. Mike's judgement, encouragement and overall guidance have been key during the production of this work, as have Dave's ready suggestions, enthusiasm and computing savoir-faire. I also thank my post-doc supervisor, Dr. Alessandro De-Vita for his kind patience during the ongoing completion of my thesis.

I dedicate this thesis to my parents, who have been a great example to me, and for all their support and inspiration.

Abstract

This thesis concerns developments and applications using the density functional theory (DFT) ab initio electronic structure method. Implementation of a pseudo atomic orbital (PAO) basis set in the linear scaling DFT program CONQUEST is reported and used to test aspects of the linear scaling algorithm. Also a separate study using plane-wave DFT (VASP code) to model the strained growth of Indium Arsenide (InAs) on the (110) surface of Gallium Arsenide (GaAs), in particular the formation of a strain relieving dislocation network, has been performed.

Pseudo atomic orbitals are the eigenstates of a pseudo-atom confined to a spherical potential, as used in the SIESTA linear scaling DFT program, and consist of a radial function multiplied by a spherical harmonic. Code to evaluate overlap and kinetic energy matrix elements between PAOs has been written, and tested using Gaussian PAOs, whose overlap integrals can be computed analytically. The PAO code has been integrated into the CONQUEST program and used to perform tests of the linear scaling algorithms on Silicon.

Conventional plane wave DFT has been applied to calculate the energetics of a dislocation network in InAs grown on GaAs(110). Both InAs and GaAs

have the zinc-blende crystal structure but the lattice constant of InAs is seven percent greater than that of GaAs. Experiments show that during deposition of the InAs by molecular beam epitaxy (MBE) compressive strain leads to formation of a strain relieving dislocation network after a critical amount of InAs coverage. In this thesis DFT is applied to calculate the energetically favoured location for the dislocation core and the resulting structure. In addition the critical InAs coverage necessary for dislocation formation is also calculated and compared to that measured by experiment.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The unifying theme of this thesis is density functional theory (DFT) which has established itself as one of the most popular theoretical approaches to the electronic structure (ES) problem, that is to find the total energy of an assembly of atoms using a quantum mechanical description of the electrons [1] [2][3]. Modern computers are able to solve the equations of quantum mechanics for many-atom systems using a variety of techniques and approximations, and the field of ES calculation has scored many successes in explaining and predicting experimental results. For example in recent years they have helped clarify semiconductor surface structures, shed light on the transition paths of chemical reactions and predict the properties of experimentally inaccessible materials (e.g. iron at the Earth's core).

The most popular approach to ES calculations used by the condensed matter physics community is density functional theory (DFT) [1][2]. The success of this approach is reflected by the award of the 1998 Nobel prize for chemistry to its theoretical founder, Walter Kohn. There are several reasons for the

present ubiquity of DFT, one of which is Bloch's theorem for electronic wavefunctions in a periodic potential, which states that they can be decomposed into a plane-wave representation. This motivated the use of plane-wave basis-sets in DFT computer codes, which had nice mathematical properties attractive to programmers interested in modelling periodic systems [4] (though of course computational DFT can be performed using many other sorts of basis-sets too). The later introduction of pseudopotentials to simplify the task of constructing electronic wavefunctions near atomic cores further increased the range of tractable systems. However the original implementations of DFT used basis functions/plane-waves which extended over the whole of the simulation cell, making the cost of calculating wavefunctions scale as the cube of the number of atoms in the cell. This algorithmic bottleneck prompted a search by groups worldwide for DFT implementations of lower cpu cost, and the local effort led by M. J. Gillan and D. R. Bowler at U.C.L. has produced a linear scaling DFT code called "CONQUEST" [5]. This thesis discusses augmenting the basis-set representation in CONQUEST, as well as applying DFT to semiconductor surfaces.

1.1 Linear Scaling DFT

Linear scaling DFT may be formulated in different ways, for example the 'divide and conquer' approach [6] splits large systems into smaller overlapping subsystems whose wavefunctions are solved separately before being recombined. Other methods are based on splitting of the density matrix itself, the Green's function approach and also density matrix minimisation, the latter of these is used in CONQUEST [7].

CONQUEST achieves linear scaling calculations on large atomic systems by enforcing localisation of the density matrix in real space. It has been shown that the elements of the density matrix decay as $1/R^d$ in metals (d is the number of spatial dimensions) and exponentially in insulators [8]. We approximate this computationally by introducing a radial cut-off in the basis functions representing the electronic wavefunctions around an atom, beyond which the basis functions (and the associated electronic densities) are zero. This leads to linear scaling calculation of the wavefunctions for large sets of atoms. Generally we refer to basis functions containing such cut-offs as "support functions" (SFs), which may themselves be composed of different types of basis functions.

Linear scaling DFT codes have been written using basis sets as varied as "spherical-wave functions" (spherical Bessel functions multiplied by a spherical harmonic), numerical functions represented on a grid, pseudo atomic orbitals (PAOs - the eigenfunctions of a pseudo-atom within a spherical box [9, 10]) and Gaussians to represent SFs. PAOs take the form of a numerical radial function multiplied by a spherical harmonic. A PAO basis is already used in the popular linear scaling code SIESTA [10], but the different architecture of CONQUEST and the possibility of designing the basis set to our own requirements convinced us to write a fresh version.

A basis of B spline (blip) [11] functions is already implemented in CONQUEST and has the important quality of being systematically convergent, so that the quality of the calculated electronic density can be improved with respect to some parameter (here grid fineness). The quality of a PAO basis set does not converge systematically, but using PAOs allows DFT energies and wavefunctions to be evaluated much more quickly than by functions

which are represented using integration grids, and though we may lose some precision we gain speed in solving for the ground-state DFT wavefunction. The usage of two basis sets allows a hierarchical scheme in which the ground state wavefunction may first be evaluated quickly using PAOs and then refined using blip functions.

1.2 Semiconductor Surfaces

There is currently a strong experimental and theoretical interest in the study of semiconductor surfaces and nanostructures, particularly in their controlled growth and fabrication. In this thesis we apply DFT methods to understand physics at semiconductor surfaces, as part of a collaboration with the experimental STM group of Professor Tim Jones at Imperial College London, who have a particular interest in the growth mechanisms of InAs on GaAs[12].

Both InAs and GaAs are III-V semiconductors with a zinc-blende structure, but the lattice constant of GaAs, 5.65 Å, is fractionally less than that of InAs (6.05 Å). Thus when InAs is deposited by molecular beam epitaxy (MBE) on the (110) surface of GaAs strain occurs, which is relieved through formation of misfit dislocation networks. This thesis describes the use of a plane-wave DFT code (VASP) to calculate the formation energies of misfit dislocations which form in InAs epilayers grown on GaAs(110) in chapter six. Here follows a summary of the chapters comprising this thesis.

1.3 Thesis Outline

In chapter two we discuss the basic theory underlying computational implementations of DFT [13]. We describe the key theorems due to Hohenberg and Kohn before giving an account of the Kohn-Sham Hamiltonian and the various terms it is composed of. We also discuss elements of the tight binding approach [14] to electronic structure due to its relevance to CONQUEST, which can obtain total energies using a number of different approximations ranging from tight binding to full ab initio DFT.

Chapter three extends the discussion into the field of linear scaling DFT [15]. In it we see why the computational expense of conventional DFT codes has tended to scale as the cube of the system size, and how this restrictive behaviour may be overcome using a reformulation of DFT in terms of the single particle density matrix which exploits its locality in electronic systems. We also discuss the specifics of how linear scaling is achieved within CONQUEST itself [5].

The author's development and implementation of computer code to calculate the matrix elements of a PAO basis set for total energy and force calculations forms the topic of chapter four. The widespread success of the SIESTA linear scaling DFT code [10], which uses PAOs, persuaded the authors of CONQUEST of the merits of such a basis, leading to its eventual incorporation into CONQUEST. In chapter four we give a thorough account of the analytic evaluation of overlap and kinetic energy integrals in terms of PAOs as well as expressions for PAO gradients, which are necessary for computing forces.

In chapter five the performance of the order N (O(N)) algorithms within CONQUEST is gauged using the new basis of PAOs. Tests are done on bulk Si and the Si(001) surface to measure the accuracy of the O(N) algorithm and its convergence towards results obtained using exact diagonalisation methods. Quantities such as the equilibrium lattice constant and bulk modulus are calculated to establish the merit of the different levels of approximation available within CONQUEST.

Application of conventional plane wave DFT (using the VASP code) to gain an understanding of misfit dislocation formation during the strained growth of InAs on GaAs [16] forms the subject of chapter six. We calculate the lowest energy misfit dislocation structure and also the InAs critical thickness (at which the strain relieving dislocations first appear) with DFT, finally comparing our results to experiment.

Chapter 2

Density Functional Theory

2.1 Introduction

This chapter serves as an introduction to density functional theory (DFT) which we use to perform electronic structure calculations throughout this thesis [1][2][3]. Density functional theory is a formalism we can apply to calculate the ground state energy of an atomic configuration, approximating the total energy of many interacting electrons with a simplification in terms of non-interacting independent particles within a mean-field (this is the Kohn-Sham (KS) reformulation). This simplification of the problem of finding the ground state energy of interacting electrons provides a tractable computational scheme which has been successfully applied to calculate many different properties of condensed matter [13][17]. Although not as accurate as methods like configuration interaction [18] (CI) or Quantum Monte Carlo (QMC), DFT allows for the treatment of larger systems of atoms because it has been made to scale as N^3 (where N is the number of atoms in the system)

rather than N^7 as for CI. For example the modelling of misfit dislocations discussed in chapter five involves up to 350 atoms in the simulation cell and such a large number would be untreatable with CI. There are cheaper computational methods available too, such as the tight binding approximation which retains the quantum mechanical nature of the bonding [14] or methods based on classical potentials but these lack the general applicability of DFT to many different materials, affording a less refined description of the interatomic bonds.

The name density functional theory derives from the fact that the total energy of the atomic system is expressed as a functional of the electron density, i.e. as an integral of various functions corresponding to the different components of the energy, with the basic variable of each function being the real space electronic density. The theory arises as the consequence of two theorems, the Hohenberg-Kohn theorems [1], the first demonstrating that the electronic energy can be expressed as a unique functional of the density, and the second stating the ground state energy is variational with respect to changes in the electronic density, so that the density which minimizes the energy (subject to physical constraints) is also the correct ground state density, in the exact theory at least. Many approximations have been necessary to turn DFT into a computational scheme, for instance the introduction of pseudopotentials (section 2.8) to replace the influence of the nuclei and core electrons on the chemically important valence electrons, or the choice of basis functions with which to represent the electronic wavefunctions. All these developments reflect attempts to reduce the expense of DFT computations whilst retaining as much predictive accuracy as possible.

In this chapter we will discuss the fundamental theorems underlying DFT

before detailing the properties of the Kohn-Sham equation (section 2.4) on which computational DFT is based. Finally we will look at the related formalism of the tight binding approach to electronic structure calculations. The Harris-Foulkes non self consistent energy functional provides a conceptual bridge between these two schemes [19]. The Harris-Foulkes functional is implemented in CONQUEST for performing quick non self consistent density functional calculations more approximate than self consistent KS approaches.

2.2 Energy as a Density Functional

To express the energy of a system of interacting electrons as a functional of the density we consider the Hamiltonian of a system of N interacting electrons. We may decompose the Hamiltonian into three terms, a term T representing the kinetic energy of the electrons, the interaction potential between electrons themselves, U, and the ion-electron potential V,

$$\begin{aligned} H &= T + U + V \\ &= H_0 + V, \end{aligned} \tag{2.1}$$

where H_0 denotes the part of the Hamiltonian associated purely with the electrons. The potential acting on the electrons due to the ions can be considered as a potential due to a static external field. We will see that the external potential plays a key role in the derivation of the two theorems due to Hohenberg and Kohn [1]. In expression 2.1 the ion-ion interactions are neglected, as they can be added in without difficulty once the electronic contributions to the Hamiltonian have been well understood. In deriving the

Hohenberg-Kohn theorems we can assume a general form for the external potential which covers the potential field due to a distribution of ionic charges. The operator V in equation 2.1 represents the action of the external field on the electrons,

$$V = \sum_{i=1}^{N_{el}} v_{ext}(\mathbf{r}_i). \tag{2.2}$$

In terms of the electronic Hamiltonian and the external potential we can write the ground state energy as [13]

$$E_0 = \langle \Psi \, | H_0 + V | \, \Psi \rangle, \tag{2.3}$$

where Ψ is the full many-body wavefunction of the system. In this formalism the expression for the electronic charge density is

$$n(\mathbf{r}) = \langle \Psi \, | \hat{n}(\mathbf{r}) | \, \Psi \rangle, \tag{2.4}$$

$$\hat{n}(\mathbf{r}) = \sum_{i=1}^{N_{el}} \delta(\mathbf{r} - \mathbf{r}_i), \tag{2.5}$$

$\hat{n}$ being the density operator. Next we shall derive the important implications of the Hohenberg-Kohn theorems relating the energy of a many electron system to the ground state charge density.

2.2.1 Hohenberg-Kohn Theorem 1

The first theorem of Hohenberg and Kohn states that for interacting electrons within an external potential $v_{ext}(\mathbf{r})$ the ground state electronic density uniquely determines $v_{ext}(\mathbf{r})$. This is very important because it tells us that given the ground state electronic density $n_0(\mathbf{r})$ all other ground state prop-

erties of the many-body system are determined by it, and hence can be expressed as a functional of $n_0(\mathbf{r})$. The HK theorems constistute an existence proof that the properties of the interacting many-body system can be expressed as a density functional, but do not tell us anything about its explicit form. It was left to KS [2] to provide an approximate functional (see next section) that could be used for practical computation of the total energy.

The proof given here follows the original route of reductio ad absurdum, where postulation of the existence of two different external potentials corresponding to the same ground state density is seen to lead to an impossible conclusion. Here we are assuming the case of the non degenerate ground state, as per the original proof [1]. We consider two different external potentials $v_{ext1}(\mathbf{r})$ and $v_{ext2}(\mathbf{r})$ differing by more than a simple additive constant. They will be associated with different Hamiltonians having different ground state wavefunctions (Ψ_1 and Ψ_2), as we are assuming a non-degenerate ground state. The variational theorem of quantum mechanics then tells us;

$$E_1 = \langle \Psi_1 | H_1 | \Psi_1 \rangle < \langle \Psi_2 | H_1 | \Psi_2 \rangle. \tag{2.6}$$

Because the difference in the two Hamiltonian operators is equal to the difference between the two external potentials v_{ext1} and v_{ext2} we can rewrite the inequality above such that

$$E_1 < E_2 + \int dr \left[v_{ext1}(\mathbf{r}) - v_{ext2}(\mathbf{r}) \right] n_0(\mathbf{r}). \tag{2.7}$$

The next step in the proof is to swap the indices of the argument in equa-

tion 2.6 and re-apply the variational principle in order to derive equation 2.8.

$$E_2 < E_1 + \int dr \left[v_{ext2}(\mathbf{r}) - v_{ext1}(\mathbf{r})\right] n_0(\mathbf{r}). \tag{2.8}$$

Adding the equations 2.7 and 2.8 leads to the contradictory statement

$$E_1 + E_2 < E_1 + E_2. \tag{2.9}$$

This cannot be true, implying that the ground state density $n_0(\mathbf{r})$ in the equations above cannot be equal for the two different external potentials, contradicting our original assumption. Thus it is proved that there is a unique correspondence between the ground state charge density of a system and the external potential. This implies the converse statement that the external potential, and therefore the entire many-body Hamiltonian, must be uniquely determined by the ground state electronic density $n_0(\mathbf{r})$.

2.2.2 Hohenberg-Kohn Theorem 2

This theorem states that the ground state energy for electrons within a given external potential can be expressed as a variational functional of the ground state density $n_0(\mathbf{r})$. The energy functional is thus minimised by the ground state density and will have a higher energy for other charge distributions. The proof of the second theorem is straightforward and relies on the Rayleigh-Ritz variational principle of quantum mechanics. As before suppose we have two different external potentials, v_{ext1} and v_{ext2} (again differing by more than an additive constant) with corresponding electronic wavefunctions Ψ_1 and Ψ_2 and Hamiltonians H_1 and H_2. We write the ground state energy, making

explicit the contribution of the external potential.

$$E_{HK}[n] = \int d\mathbf{r} v_{ext}(\mathbf{r}) n(\mathbf{r}) + F_{HK}[n]. \tag{2.10}$$

For a system having a ground state density $n_1(\mathbf{r})$ and a corresponding external potential $v_{ext1}(\mathbf{r})$ the HK functional will be equal to the ground state expectation value of the Hamiltonian,

$$E_1 = E_{HK}[n_1] = \langle \Psi_1 | \hat{H}_1 | \Psi_1 \rangle. \tag{2.11}$$

However, if we consider some other density $n_2(\mathbf{r})$ which corresponds to a different external potential the energy will be higher than for n_1 due to the variational principle of quantum mechanics,

$$E_1 = \langle \Psi_1 | \hat{H}_1 | \Psi_1 \rangle < \langle \Psi_2 | \hat{H}_1 | \Psi_2 \rangle = E_2. \tag{2.12}$$

Hence the energy of the universal functional written in equation 2.10 will have a global minimum at the ground state charge density. The derivations so far have relied on a real space representation of the wavefunctions and operators, as this allows us to see the locality in real space of the quantities comprising the total energy. Later we will see that the introduction of the pseudopotential approximation (2.8) disrupts this formal locality somewhat.

2.3 The Kohn-Sham Density Functional

Though the HK theorems tell us that there exists a universal functional for the energy of an interacting many electron system they do not say anything

about its explicit form. In 1965 Kohn and Sham published a paper outlining an energy functional [2] which mapped the interacting many electron problem onto an auxiliary system of non-interacting fictitious particles.

As the true many-body functional will have a global minimum at the ground state density we can write down the condition for it to be stationary with respect to small changes in the density (this is the Euler equation for the functional [20]). Though we do not know the specific form of the functional we may separate F[n] into two parts, T[n] and G[n]. T[n] refers to the kinetic energy of non-interacting electrons and all other electronic contributions to the energy are bracketed into G[n]. Importantly we will see that the Euler equation, expressing the condition that the total energy is stationary with respect to fluctuations in the density, has the same form whether the electrons are interacting or not. Using the calculus of variations [20] we may express a small variation of the total electronic energy as

$$\delta E = \int d\mathbf{r}[V(\mathbf{r}) + \frac{\delta T}{\delta n(\mathbf{r})} + \frac{\delta G}{\delta n(\mathbf{r})}]\delta n(\mathbf{r}). \tag{2.13}$$

Setting the terms within the square brackets equal to zero ensures that δE becomes zero too, but we must also take into account an important constraint which is that the total number of electrons must remain fixed. We can add this constraint into the Euler equation as a Lagrange multiplier, obtaining equation 2.14 where μ, the chemical potential appears as the Lagrange multiplier.

$$\frac{\delta T}{\delta n(\mathbf{r})} + V(\mathbf{r}) + \frac{\delta G}{\delta n(\mathbf{r})} = \mu. \tag{2.14}$$

Significantly equation 2.14 also has the same form for non-interacting electrons. In the case of non-interacting electrons G = 0 as only the individual kinetic energies and interaction with the external potential will contribute to

the energy. As such the Euler equation becomes

$$\frac{\delta T}{\delta n(\mathbf{r})} + V(\mathbf{r}) = \mu. \tag{2.15}$$

We can rewrite the Euler equation of the interacting system as

$$\frac{\delta T}{\delta n(\mathbf{r})} + V_{eff}(\mathbf{r}) = \mu \tag{2.16}$$

where we have set

$$V_{eff}(\mathbf{r}) = V(\mathbf{r}) + \frac{\delta G(\mathbf{r})}{\delta n(\mathbf{r})}. \tag{2.17}$$

KS set the external potential in the Euler equation of the non-interacting electrons equal to the last two terms of equation 2.14, in doing so they turned the HK theorems into a tractable computational scheme, although we have as yet no explicit form for the derivative of G with respect to n(r). We can find the ground state density of the non interacting electrons through a modified Schrodinger equation, now with potential $V_{eff}(r)$, by solving it for the eigenfunctions of the Hamiltonian,

$$\frac{-\hbar^2}{2m}\nabla^2\psi_i + V_{eff}\psi_i = \epsilon_i\psi_i. \tag{2.18}$$

This Hamiltonian is known as the Kohn-Sham (KS) Hamiltonian. Once we have obtained the eigenfunctions we can find the ground state density through relationship 2.19. The individual eigenfunctions of the KS Hamiltonian have no physical meaning since they relate to fictitious independent particles, but the charge density of the interacting electrons can be derived from them;

$$n(\mathbf{r}) = \sum_{i=1}^{N} |\psi_i(\mathbf{r})|^2. \tag{2.19}$$

Using the charge density we can then derive key properties of the ground state such as the total energy and the interatomic forces.

So far we have not required an explicit form for the effective potential which enters the KS Hamiltonian, but it becomes necessary for purposes of computation. As well as introducing the simplification in terms of non-interacting particles KS also proposed an explicit form for the effective potential forming the mean field. They divided it into three separate terms, $v_{ext}(\mathbf{r})$ representing the potential of the ionic cores, $V_H(\mathbf{r})$ the Hartree potential which is the classical Coulomb energy of the self-interacting charge density, and the mysterious exchange-correlation potential V_{xc} whose purpose was to include all quantum mechanical contributions to the energy neglected by the remainder of the KS Hamiltonian of equation 2.18 with V_{eff} as defined by equation 2.20 below.

$$V_{eff}(\mathbf{r}) = v_{ext}(\mathbf{r}) + V_H(\mathbf{r}) + V_{xc}(\mathbf{r}). \tag{2.20}$$

In the following sections we will look at each of these terms in more detail, establishing their explicit dependence on the density $n(\mathbf{r})$ in order to make explicit the KS Hamiltonian.

2.4 Kohn-Sham Effective Potential

Having established that the energy of a many-electron system can be expressed as a universal functional of the ground state density, KS proceeded to suggest an approximate density functional based on the simplified case of the homogeneous electron gas. The derived functional can then be applied to electronic systems with a non-uniform density distribution, using what is

known as the local density approximation for the exchange-correlation, giving surprisingly accurate energies considering the nature of the approximation. Next we look at the Hartree term in the effective potential, before discussing the more complex exchange-correlation potential.

2.4.1 Hartree Energy

We have rewritten the total energy functional in terms of the kinetic energy of non-interacting electrons having the same density, and an additional part G[n];

$$F[n] = T[n] + G[n]. \tag{2.21}$$

Now G[n] will be the sum of the Hartree and exchange-correlation energy functionals. The Hartree energy is the electrostatic Coulomb energy for the same density distribution of static electrons,

$$E_H = \frac{1}{2}e^2 \int d\mathbf{r}d\mathbf{r}' \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.22}$$

The exchange-correlation term in G[n] is actually defined to be the difference between the energy of the non-interacting system expressed in the rest of the KS functional and the total energy of equivalent interacting electrons. It plays the role of restoring the quantum mechanical interplay of the many electron system as a functional of the independent particle density.

$$E_{tot}[n] = \int d\mathbf{r} v_{ext}(\mathbf{r})n(\mathbf{r}) + T[n] + E_H[n] + E_{xc}[n]. \tag{2.23}$$

The separation of G[n] into $E_H[n]$ and $E_{xc}[n]$, the Hartree and exchange-correlation energy functionals is a crucial step in formulating the KS total

energy expression. Now explicit in $n(\mathbf{r})$, equation 2.23 provides a starting point for derivation of the stationary condition (Euler equation) giving the ground state energy of the system subject to the constraints that $v_{ext}(\mathbf{r})$ and the total electron number remain fixed during calculation.

2.4.2 Exchange Correlation Functional

No general analytic form for the exchange-correlation energy functional is known but successful approximations have been made. In their paper [2] KS derive an energy functional for the case of the homogeneous electron gas, where they found that the effects of exchange and correlation could be approximated by a function local in the charge density. In what is now called the local density approximation (LDA) they set the exchange correlation energy functional in a solid equal to the functional of a uniform electron gas of the same charge density. Although the exchange energy of the uniform gas is known analytically the correlation energy is not, it has instead to be estimated numerically and then fitted to a simple parameterised form for inclusion in the KS energy functional.

The expression for the exchange energy per particle in the homogeneous electron gas is that derived in Hartree-Fock theory [13], the expression for it is given in equation 2.24 below, it is clearly a function local in the density $n(\mathbf{r})$.

$$\epsilon_x = -\frac{3}{4}\left(\frac{6}{\pi}n\right)^{\frac{1}{3}}. \tag{2.24}$$

Taking the exchange-correlation energy per electron to be $\epsilon_{xc}(n(\mathbf{r}))$ in the uniform gas we can then write down an estimate for the exchange-correlation

energy of a non-uniform general electron gas,

$$E_{xc} \approx \int d\mathbf{r} n(\mathbf{r}) \epsilon_{xc}(n(\mathbf{r})). \tag{2.25}$$

As already mentioned expression for the correlation energy is not a known quantity and must be made by fitting to numerical simulations of the electron gas to an analytic form. A popular choice of fit is to the Ceperley-Alder Monte Carlo simulations of the correlation energy of the uniform electron gas [21].

The substantial success of the LDA in predicting condensed matter properties led to development of further functional approximations for the exchange-correlation energy incorporating the gradient of the charge density as well, in so called GGA functionals (this concept was also outlined in the original KS paper [2]). Many different analytic forms have been used for the fitting procedure and a few of these are discussed in reference [13]. Once the form of the exchange correlation energy term is established the corresponding potential can be expressed as

$$V_{xc}(\mathbf{r}) = \left(\epsilon_{xc} + n \frac{\partial \epsilon_{xc}}{\partial n} - \nabla (n \frac{\partial \epsilon_{xc}}{\partial \nabla n}) \right)_{\mathbf{r}}. \tag{2.26}$$

Though the LDA often gives good predictions of elastic and vibrational properties of solids it tends to underestimate the bulk lattice constants of crystals whilst overestimating cohesive energies [13]. Use of the GGA functional is found to provide better cohesive energy estimates without the overbinding predicted by LDA, though it tends to overcorrect the lattice parameter estimates, giving values larger than experiment.

Even though a large approximation has been made in using the exchange-correlation functional of a uniform gas to describe a system of varying density, both LDA and GGA provide good predictions of trends and energies. There are physical reasons for the unexpected accuracy of the local density approximation in describing non-uniform electronic configurations. Importantly it preserves the sum rule that the exchange-correlation hole around an electron integrates to one. The hole itself is a consequence of exchange-correlation effects preventing fermions from occupying the same quantum states, and the combination of the electron and associated hole has zero charge. Though LDA may not predict the correct form of the hole its satisfaction of the sum-rule provides results which compare well with experiment.

2.5 The Effective Potential

Having discussed the Hartree and exchange-correlation contributions to the total energy we can now derive the effective potential as an explicit functional of the charge density. The effective potential is the derivative of the energy with respect to the electron density, returning to equation 2.21

$$\begin{aligned} V_{eff}(\mathbf{r}) &= v_{ext}(\mathbf{r}) + \frac{\delta G}{\delta n(\mathbf{r})} = v_{ext}(\mathbf{r}) + \frac{\delta E_H}{\delta n(\mathbf{r})} + \frac{\delta E_{xc}}{\delta n(\mathbf{r})}. \\ &= v_{ext}(\mathbf{r}) + V_H(\mathbf{r}) + V_{xc}(\mathbf{r}). \end{aligned} \tag{2.27}$$

Substituting in the explicit form of the Hartree energy 2.22 provides us with the following expression,

$$V_H(\mathbf{r}) = e^2 \int d\mathbf{r}' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.28}$$

We can also write the explicit form for the exchange correlation potential,

$$V_{xc} = \frac{\delta}{\delta n(\mathbf{r})} \int d\mathbf{r}_1 n(\mathbf{r}_1) \epsilon_{xc}(n(\mathbf{r}_1)) = \mu_{xc}(n(\mathbf{r})). \tag{2.29}$$

with

$$\mu_{xc}(n) = \frac{d}{dn}(n\epsilon_{xc}(n)). \tag{2.30}$$

Thus we can write the Kohn-Sham effective potential as

$$V_{eff}(\mathbf{r}) = v_{ext}(\mathbf{r}) + e^2 \int d\mathbf{r}' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + \mu_{xc}(n(\mathbf{r})). \tag{2.31}$$

Having established the effective potential as a functional of the electronic density we can now solve the KS Hamiltonian for the total energy. There are two broad approaches to its solution, the revolutionary step made by Car and Parrinello [22] involved a direct minimisation of the energy functional through a simulated annealing procedure. Alternative approaches rely on successive refinements of trial input densities $n_{trial}(\mathbf{r})$ till the correct ground state density $n_{gs}(\mathbf{r})$ is found. For example an initial (trial) charge density may be guessed at (by summing up the charge densities of the individual atoms) and used to form the KS effective potential. Then the resulting KS Hamiltonian may be solved for the eigenorbitals of the non interacting electrons, which can be used to find a new electron density.

$$\left[-\frac{1}{2}\nabla^2 + V_{eff}(\mathbf{r})\right] \psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}). \tag{2.32}$$

If the electron density obtained from the KS equation is identical to that used to form the effective potential a self consistent solution has been obtained, and there is no need to form another input trial density. If it is not then another trial density can be formed, usually through a judicious combination

of previous input densities, until the self consistent solution is found.

2.6 The Harris-Foulkes Functional

Until now we have presented a single expression for the Kohn-Sham energy functional which gives the correct variational ground state energy of the system. However different functionals may be found through transformations of the original equation which still yield the correct ground state energy for the ground state charge density. An important example of this is the Harris-Foulkes (HF) non self-consistent functional for the energy [19], which returns a total energy as a functional of the input trial density only, without the need to form a fresh output density. The HF functional is important within CONQUEST, for reasons which will be outlined below.

In order to write down the expression for the HF energy functional we follow [13], establishing an expression for the band structure energy in which a sum over the KS eigenvalues is used in writing down the HF functional. The KS eigenvalues can be written as

$$\epsilon_i = \langle \psi_i | H_{KS} | \psi_i \rangle. \tag{2.33}$$

Thus we can rewrite the (non interacting) kinetic energy in terms of the sum over KS eigenvalues minus an integral involving the KS effective potential,

$$T[n] = E_s - \int d\mathbf{r} V_{in}(\mathbf{r}) n_{out}(\mathbf{r}). \tag{2.34}$$

Here V_{in} is the KS effective potential constructed using the guessed input

charge density. The quantity E_s, the band energy, represents the sum over eigenvalues,

$$E_s = \Sigma_{i=1}^{N_{el}} \epsilon_i. \tag{2.35}$$

Importantly E_s is also a functional of the density, being the ground state energy of the non interacting electron system. Writing the band energy explicitly in terms of the potentials we have introduced so far,

$$E_s = T[n] + \int d\mathbf{r} n(\mathbf{r}) v_{ext} + \int d\mathbf{r} n(\mathbf{r}) V_H(\mathbf{r}) + \int d\mathbf{r} n(\mathbf{r}) V_{xc}(\mathbf{r}). \tag{2.36}$$

We can rewrite the last two terms of the right hand side

$$\int d\mathbf{r} n(\mathbf{r}) V_H(\mathbf{r}) = e^2 \int d\mathbf{r} d\mathbf{r}' \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.37}$$

$$\int d\mathbf{r} n(\mathbf{r}) V_{xc}(\mathbf{r}) = \int d\mathbf{r} n(\mathbf{r}) \mu_{xc}(n(\mathbf{r})). \tag{2.38}$$

Though the expressions 2.37 and 2.38 are related to the Hartree and exchange-correlation energies they are not equivalent, thus to write the KS energy in terms of the band energy we must also add terms correcting for the difference in the Hartree and exchange-correlation energies respectively. These are called the "double-counting correction" terms,

$$E_{tot} = E_s + \Delta E_H + \Delta E_{xc} + E_{II}, \tag{2.39}$$

E_s is the band structure energy as before, $\Delta E_H + \Delta E_{xc}$ are the double counting corrected Hartree and exchange-correlation energies,

$$\begin{aligned} \Delta E_H &= -\frac{1}{2} \int d\mathbf{r} n_{in}(\mathbf{r}) V_H(\mathbf{r}) \\ \Delta E_{xc} &= \int d\mathbf{r} n_{in}(\mathbf{r}) (\epsilon_{xc}[n_{in}(\mathbf{r})] - \mu_{xc}[n_{in}(\mathbf{r})]). \end{aligned} \tag{2.40}$$

E_{II} is the interaction energy of the ionic cores.

The HF energy functional provides a good estimate for the true ground state energy at densities which are close to the self consistent solution. Crucially the error in the HF energy and the KS energy is related to the square of the error in the guessed input charge density. Thus for input charge densities which are close to the ground state solution the HF energy provides a very good estimate of the true ground state energy. Thus it may be used for speedy evaluation of energies and forces, for example using the sum of spherically symmetric atomic charge densities to form the input. The difference between the HF energy and the KS energy for a given input density is also a useful measure of the lack of self-consistency during an iterative search for the ground state.

2.7 Forces from the Kohn-Sham Functional

Thus far the ionic cores have entered the picture through the external potential $v_{ext}(\mathbf{r})$. This contribution to the electronic energy is given by integrating its product with the electronic density over all space,

$$E_{ext} = \int d\mathbf{r} v_{ext}(\mathbf{r}) n(\mathbf{r}). \tag{2.41}$$

From the KS equations we can obtain the ground state charge density of elecrons within a fixed external potential. However if we wish to find the lowest energy configuration for an assembly of atoms we must also alter the ionic positions as well as the electronic orbitals. To do this we must find an expression for the force which the electrons exert on the ions. The lowest

energy configuration for our set of atoms will then be found by successively relaxing the electronic orbitals and ionic positions until the interatomic forces are deemed negligible.

The force acting on a nucleus within our ensemble of atoms is given by the gradient of the total energy with respect to nuclear position,

$$f_i = -\nabla_{\mathbf{R}_i} E_{tot}. \tag{2.42}$$

The force due to other nuclei is relatively simple being a sum of the standard Coulomb force over all the other nuclei. In addition to this we must also find the force exerted by the surrounding electrons, so we return to the Kohn-Sham expression for the electronic energy,

$$E_{ks} = \int d\mathbf{r} v_{ext}(\mathbf{r}) n(\mathbf{r}) + F[n(\mathbf{r})]. \tag{2.43}$$

Now the external potential $v_{ext}(\mathbf{r})$ is given by a sum over the individual ionic potentials present in the system. Writing down an expression for the gradient of the Kohn-Sham electronic energy with respect to the ith ionic position,

$$\begin{aligned} -\nabla_i E_{ks} &= -\int d\mathbf{r} n(\mathbf{r}) \nabla_{\mathbf{R}_i} v_{ext}(\mathbf{r}) \\ &- \int d\mathbf{r} v_{ext}(\mathbf{r}) \nabla_{\mathbf{R}_i} n(\mathbf{r}) \\ &- \int d\mathbf{r} \frac{\delta F}{\delta n(\mathbf{r})} \nabla_{\mathbf{R}_i} n(\mathbf{r}). \end{aligned} \tag{2.44}$$

The last term in equation 2.44 has been rearranged in terms of $\nabla_{\mathbf{R}_i} n(\mathbf{r})$ for a purpose, recalling that the Kohn-Sham functional is stationary at the electronic ground state as in equation 2.16 then the last two terms can be

rewritten

$$\int d\mathbf{r} \left(v_{ext}(\mathbf{r}) + \frac{\delta F}{\delta n(\mathbf{r})} \right) \nabla_{\mathbf{R}_i} n(\mathbf{r}) = \int d\mathbf{r} \mu \nabla_{\mathbf{R}_i} n(\mathbf{r}), \tag{2.45}$$

which is equal to zero because the total electron number is not variable with respect to nuclear positions. This derivation is related to the Hellmann-Feynman theorem which states that the force acting on a given nucleus i may be written

$$\frac{\partial E}{\partial \mathbf{R}_i} = \langle \Psi | \frac{\partial H}{\partial \mathbf{R}_i} | \Psi \rangle. \tag{2.46}$$

This expression is derived from the starting point,

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{R}_i} &= \frac{\partial}{\partial \mathbf{R}_i} \langle \Psi | H | \Psi \rangle \\ &= \langle \frac{\partial \Psi}{\partial \mathbf{R}_i} | H | \Psi \rangle + \langle \Psi | \frac{\partial H}{\partial \mathbf{R}_i} | \Psi \rangle + \langle \Psi | H | \frac{\partial \Psi}{\partial \mathbf{R}_i} \rangle \\ &= \langle \Psi | \frac{\partial H}{\partial \mathbf{R}_i} | \Psi \rangle + E \frac{\partial}{\partial \mathbf{R}_i} \langle \Psi | \Psi \rangle \\ &= \langle \Psi | \frac{\partial H}{\partial \mathbf{R}_i} | \Psi \rangle \end{aligned} \tag{2.47}$$

again since the derivative of the total electronic density with respect to ionic coordinates is zero. Thus the electronic contribution to the force on the ions will be

$$f_{elec} = - \int d\mathbf{r} n(\mathbf{r}) \nabla_{\mathbf{R}_i} v_{ext}(\mathbf{r} - \mathbf{R}_i), \tag{2.48}$$

as the gradient operator is non zero only on the ith component of the sum over ionic potentials $v(\mathbf{r})$. Thus the total force on the ions is the sum of the Coulomb force between the ions themselves and the force in equation 2.48 due to the electrons.

2.8 Pseudopotentials

Although DFT provides a reasonable calculational framework there are many practical approximations required to make computation feasible. Electronic wavefunctions close to an atomic core oscillate far more rapidly (on the scale of hundredths of an angstrom) than they do further away from the core. Yet most of the chemical properties of materials depend on the valence electrons, which interact only weakly with the atomic nucleus in comparison to the core electrons, which "screen" the valence electrons from the strong attraction of the nucleus. The cost of simulating the rapidly oscillating wavefunctions of core electrons is much larger than that for valence electrons as many more degrees of freedom are required in the basis set. The pseudopotential is a device which allows us to neglect the expensive explicit treatment of core electrons and focus our calculations on the valence electrons, approximating the behaviour of the nucleus and core electrons together within an artificial potential constrained to reproduce basic characteristics of the interaction between the ionic core and the valence electrons.

The action of core orbitals on the valence electrons contributes to the electrostatic and exchange-correlation terms, and the wavefunctions of the valence electrons must also be orthogonal to those of the core in accordance with the Pauli exclusion principle. If the effects of valence-core orthogonality were not enforced during a calculation the valence orbitals could simply remain close to the atomic core in order to lower their energies. We can reflect this orthogonality through use of a modified core potential including the effects of the nucleus and the core electrons together. The modified repulsive potential, or pseudopotential, prevents the valence electrons from becoming core electrons and also removes computationally expensive oscillations of valence

electrons close to the core.

2.9 Defining the Pseudopotential

The "ionic pseudopotential" is limited to a spherical region surrounding the atomic centre, and should preserve the energies of the valence orbitals. The pseudopotential will be dependent on the angular momentum l, as well as m, the magnetic quantum number. The dependence on l makes the pseudopotential a non-local operator. Thus the total pseudopotential operator must be expressed as a sum over the individual angular momentum channels leading to a "semi-local" form (it is called semi-local because there is no requirement for different functions to represent different radial regions),

$$\hat{V}_{ps} = \sum_{lm} |Y_{lm}\rangle V_l(r) \langle Y_{lm}| \tag{2.49}$$

where the functions Y are spherical harmonics. The contribution to the KS effective potential is made by evaluating the matrix elements of $\hat{V}_{ps}$ between the KS eigenfunctions. The pseudopotential operator can be expressed as the sum of a local and an l dependent part,

$$V_{ps}(\mathbf{r}) = V_{loc}(\mathbf{r}) + \delta V_l(\mathbf{r}) \tag{2.50}$$

2.10 Norm Conservation

The norm-conservation condition makes pseudopotentials more accurate and transferable (approximating with consistency the effect of the ionic core

within different chemical environments). It is based on the principle that the pseudopotential should preserve not only the properties of the valence eigenstates but also the properties of eigenstates which are close in energy to these. This is important for reproducing electronic interactions between atoms, since when they are brought close together the valence energy states form band states which may cover an energy range of tens of electron volts.

The principle of norm conservation can be put into a simple form, let ψ_{ae} be an all electron valence eigenfunction and ψ_{ps} be the corresponding pseudopotential eigenfunction then we can write

$$\int_0^{R_c} d\mathbf{r}\mathbf{r}^2\psi_{ae}^2(\mathbf{r}) = \int_0^{R_c} d\mathbf{r}\mathbf{r}^2\psi_{ps}^2(\mathbf{r}). \tag{2.51}$$

The equation 2.51 gives us a clue why as to the name of the principle, it is the condition that the charge contained by ψ_{ae} within the core region should be equal to that contained by ψ_{ps} within the same. Hamann, Schluter and Chiang [23] showed that the norm conservation condition of equation 2.51 also guaranteed that the first energy derivative of the logarithmic derivatives of ψ_{ps} and ψ_{ae} agree at and beyond the core radius, R_c.

To understand the importance of norm-conservation we can consider a valence eigenstate lying within the range we are concerned with. We can find the valence wavefunction by integrating the Schrodinger equation outwards from the origin to a point beyond the core radius R_c. We would like the valence wavefunction beyond R_c to be the same whether we use the all electron potential or the pseudopotential. In fact if $\psi(r)$ and its first derivative are preserved at R_c it follows that the two quantities will be in agreement throughout the valence region as the two potentials are identical beyond this

point. We can reduce this condition to the requirement that $(d\psi/dr)/\psi = f$ (the logarithmic derivative of ψ) should match outside the core radius. Norm conservation is equivalent to the requirement that df/dE should match outside the core radius [23], and provides a convenient way of ensuring this.

2.11 Kleinman-Bylander Formulation

The Kleinman-Bylander formulation of pseudopotentials uses a representation in terms of radial functions multiplied by spherical harmonics, which is the same functional form as the PAO basis functions whose implementation we describe later, hence we discuss some of the details here. Kleinman and Bylander [24] showed that the pseudopotential operator $V(\mathbf{r}, \mathbf{r}')$ could be expressed in a separable form, that is a sum over products of individual functions of $\mathbf{r}$ and $\mathbf{r}'$, $\sum_i f_i(\mathbf{r})g_i(\mathbf{r}')$

$$\hat{V}_{KB} = V(\mathbf{r}) + \sum_{lm} \frac{|\psi_{lm}^{PS}\delta V_l\rangle\langle\delta V_l\psi_{lm}^{PS}|}{\langle\psi_{lm}^{PS}|\delta V_l|\psi_{lm}^{PS}\rangle}. \tag{2.52}$$

The pseudopotential operator has been decomposed into local and non local components following equation 2.50. Now the second term in equation 2.52 above is fully separable in each of the three spherical polar coordinates. In the equation above we have the Kleinman-Bylander projectors which project onto the wavefunction as follows,

$$\langle\delta V_l\psi_{lm}|\psi\rangle = \int d\mathbf{r}\delta V_l(\mathbf{r})\psi_{lm}^{PS}(\mathbf{r})\psi(\mathbf{r}). \tag{2.53}$$

The KB separable form for the pseudopotential operator is advantageous because the matrix elements of the pseudopotential operator can be formed

using products of the projection operations defined in equation 2.53. The matrix elements are obtained by

$$\langle\psi_i|\delta V_{NL}|\psi_j\rangle = \sum_{lm}\langle\psi_i|\psi_{lm}^{PS}\delta V_l\rangle\frac{1}{\langle\psi_{lm}^{PS}|\delta V_l|\psi_{lm}^{PS}\rangle}\langle\delta V_l\psi_{lm}^{PS}|\psi_j\rangle. \tag{2.54}$$

The form of the functions ψ_{lm}^{PS} is

$$\psi_{lm}^{PS} = \psi_{lm}(r)Y_{lm}(\Omega). \tag{2.55}$$

Thus the functional form of the non local pseudopotential projectors and the PAO functions is identical and the same kernel of code which is used to evaluate matrix elements between PAOs themselves can also be used to evaluate the non local KB pseudopotential matrix elements.

2.12 Tight binding Theory

Here we describe the tight binding approach to electronic structure calculations as this is an approximation that can be used in CONQUEST with the newly implemented PAO basis. In tight binding theory the electronic orbitals of a group of atoms are expressed as linear combinations of the orbitals of the individual atoms themselves [14]. Thus the wavefunction for the system can be written as a linear combination of atomic orbitals,

$$\psi(\mathbf{r}) = \sum_{i\alpha} C_{i\alpha}\phi_{i\alpha}(\mathbf{r}). \tag{2.56}$$

The index i refers to the atom on which the orbitals, referenced by α, are based. The atomic-like orbitals may for example be pseudo-atomic orbitals

as used in the tight-binding code of Sankey and Niklewski [9] and later in the linear scaling density functional theory code SIESTA [10], and now also in the CONQUEST code which we will discuss later in this thesis.

An important paper in the history of the tight binding approach was that of Slater and Koster (SK) [25], who developed a parameterized form for the matrix elements of the Hamiltonian - the essence of the tight binding approximation. In order to develop an explicit expression for the tight binding Hamiltonian SK developed basis functions which were Bloch sums over Lowdin orbitals. A Bloch sum runs over all of the periodic images of an atomic orbital $\phi_{i\alpha}(\mathbf{r}-\mathbf{R}_i)$, returning a basis state with wavevector $\mathbf{k}$.

$$\chi_{\mathbf{k}\alpha}(\mathbf{r}) = N^{-\frac{1}{2}} \sum_{\mathbf{R}_i} \exp[i\mathbf{k}\cdot\mathbf{R}_i]\phi_{i\alpha}(\mathbf{r}-\mathbf{R}_i). \tag{2.57}$$

Here N is an infinite number of periodic images [14]. Atomic orbitals on different ions will not define an orthogonal set of functions, to simplify the analysis the transformation of Löwdin is used to convert the set of non orthogonal orbitals into an orthogonal one,

$$\psi_{i\alpha} = \sum_{i'\alpha'} S^{-\frac{1}{2}}_{i\alpha i'\alpha'} \phi_{i'\alpha'}. \tag{2.58}$$

The cost of forming an orthogonal basis is that the Löwdin functions will have a greater spatial extent than the atomic orbitals from which they were formed. The Hamiltonian expressed using Löwdin functions may thus be non negligible even between third nearest neighbour atoms. SK formed Bloch sums over the Löwdin functions to create a basis set for the Hamiltonian

matrix,

$$H_{i\alpha j\beta} = N^{-1} \sum_{\mathbf{R}_i \mathbf{R}_j} \exp[i\mathbf{k}\cdot(\mathbf{R}_j - \mathbf{R}_i)] \times \int \psi^*_{i\alpha}(\mathbf{r}-\mathbf{R}_i) H \psi_{j\beta}(\mathbf{r}-\mathbf{R}_j) d\mathbf{r}. \quad (2.59)$$

Importantly the sum over periodic images exactly cancels the normalistion factor N, thus the Hamiltonian can in fact be expressed as a single sum over the periodic images of atomic sites,

$$H_{i\alpha j\beta} = \sum_{\mathbf{R}_j} \exp[i\mathbf{k}\cdot(\mathbf{R}_j - \mathbf{R}_i)] \times \int \psi^*_{i\alpha}(\mathbf{r}-\mathbf{R}_i) H \psi_{j\beta}(\mathbf{r}-\mathbf{R}_j) d\mathbf{r}. \quad (2.60)$$

The step made by SK was to avoid explicit evaluation of the integral in the expression 2.60 by substituting a parameterized form for the Hamiltonian matrix elements, depending only on the distance $|\mathbf{R}_i - \mathbf{R}_j|$. Ordinarily integrals such as the one in equation 2.60 can involve up to three atomic centres, for example when the two orbitals and the potential part of the Hamiltonian operator all lie on different atoms. However SK made the two-center approximation, restricting the potential terms in the Hamiltonian to the atoms i and j at which the orbitals are sited. Thus the Hamiltonian matrix elements can be written more simply as

$$H_{i\alpha j\beta} = \sum_{\mathbf{R}_j, J} \exp[i\mathbf{k}\cdot(\mathbf{R}_j - \mathbf{R}_i)] h_{\alpha\beta J}(|\mathbf{R}_j - \mathbf{R}_i|) G_{\alpha\beta J}(k,l,m). \quad (2.61)$$

The total angular momentum of the bond is J, l and m are the conventional orbital and azimuthal angular momentum quantum numbers. $G_{\alpha\beta J}$ is the angular term in the integral, as specified in [25].

Although the above formalism is sufficient for discovering the band energy of the electrons it does not describe the total energy of an atomic system.

Chadi [26] added in a pairwise repulsive energy,

$$\begin{aligned} E_{tot} &= E_{band} + E_{rep} \\ E_{rep} &= \sum_{ij} U_{ij}. \end{aligned} \tag{2.62}$$

which improved the applicability of the tight binding formalism to electronic structure calculations [14].

The tight binding approach to electronic structure calculations is more approximate than the method of density functional theory. Yet there are connections between the two formalisms, discussed in much depth by Foulkes and Haydock [19], and Sutton et al [27] who present an argument for the validity of the pairwise repulsive potential in the tight binding energy functional by analysing the properties of the KS functional. The original development of tight binding theory was related to the band structure of periodic solids and made use of reciprocal space representation of the electronic wavefunctions [13]. The recent trend towards electronic structure schemes which scale linearly with the number of atoms in the system has placed more emphasis on the real space represention of wavefunctions and charge densities.

CONQUEST is capable of running density functional theory calculations from tight-binding levels of accuracy up to full self-consistent solution of the Kohn-Sham functional, for example using the Harris-Foulkes energy functional to perform non self-consistent ab initio tight binding calculations. We shall see in the next chapter that the hierarchy of approximations available within CONQUEST rely on enabling the variation of separate parameter sets during the search for the ground state.

Chapter 3

Linear Scaling DFT

3.1 Introduction

As the computational expense of plane wave DFT implementations has begun to stabilise workers in the field have striven for new solutions of the KS equations with better scaling than the N_{atoms}^3 of existing codes. Justifying such new approaches are arguments that the distribution of the electron density around an atom depends most strongly on its local charge environment.

In this chapter we will first discuss the scaling of conventional approaches to DFT described in the previous chapter (section 3.2), before outlining results on the locality of Wannier functions (section 3.4) and the density matrix (section 3.5) in electronic systems. We then describe the density matrix minimisation approach to linear scaling DFT in section 3.7. Finally we will take a look at the specific scheme (minimisation with respect to both the density matrix and the basis functions on which it is expressed) used within the

CONQUEST code, discussing details of its implementation and performance in section 3.8 until the end of the chapter.

3.2 Scaling Trouble

There are two main factors which constrain the maximum size of systems treatable with DFT, namely software and hardware. Despite the seemingly inexorable increase of modern cpu power, inefficiencies in algorithmic implementation can drastically curb the scope of a computational method. The breakthrough of the Car-Parrinello paper [22] was in establishing a highly innovative implementation of computational DFT with ground breaking efficiency rather than proving new identities or theorems.

Therefore crucial to the successes of computational condensed matter are the scaling behaviours of calculations with respect to the system size. The current popularity of DFT is largely due to its N^3 scaling (where N is the number of atoms) within the context of other electronic structure approaches which, despite offering more accuracy, have much greater computational expense. The configuration interaction (CI) method for example goes as N^7 [18], and quantum monte carlo, despite being made to scale as N^3, suffers from a much greater computational prefactor which makes treating large systems challenging [28]. More recently the realisation that DFT can be implemented in a linear scaling fashion [29] has led to development of codes such as CONQUEST [5], ONETEP [30] and SIESTA [10] which exploit the localisation of the single particle density matrix in solving the KS Hamiltonian. In fact the success of the pseudo atomc orbital (PAO) based SIESTA code is of particular relevance to this thesis, a large part of which is concerned with the

implementation of PAOs within CONQUEST.

Originally DFT methods relied on direct diagonalisation of the Hamiltonian matrix in order to find its eigenstates and eigenvalues. The cost of diagonalising a square matrix of size L is L^3 as explained in [31], the size of the hamiltonian being equal to the number of basis functions used to express it. This scaling was reduced by the advent of new methods to solve the KS equations like the conjugate gradients schemes discussed in the previous chapter [32] or the Car-Parrinello approach [22]. The new methods scaled as $L\ ln(L)$ when using a plane wave basis rather than as the L^3 of the direct diagonalisation schemes. However, notwithstanding the improved scaling with respect to the basis set, the new methods still scaled as N^3_{atoms} with respect to the total number of atoms. The scaling derives from the orthogonality constraint of the KS orbitals during a calculation, for example consider the following expression, which must be evaluated in order to enforce orthogonality,

$$Q_{mn} = \int d\mathbf{r}\psi^*_m(\mathbf{r})\psi_n(\mathbf{r}) \tag{3.1}$$

$$Q_{mn} \approx \sum_l \delta(r)\psi^*_m(\mathbf{r}_l)\psi_n(\mathbf{r}_l). \tag{3.2}$$

The second equation 3.2 refers to integration on a grid which is used in practical evaluation of expression 3.1. There are three indices which grow with system size in equation 3.2, l is the number of integration grid points, and n,m are indices labelling the Kohn Sham orbitals. Evaluating the integral for all values of l, m and n is thus a computation which scales with the cube of the system size.

3.3 Nearsightedness

Historically solid state physics has been framed in terms of extended Hamiltonian eigenstates of periodic crystals, characterised by wavevectors k from Bloch's theorem. In many body systems the electronic wavefunctions are dependent on each other throughout all space, for example the Pauli exclusion principle applies to electronic orbitals no matter how far apart they might be. Accurate description of phenomena such as the Fermi surface in k-space and critical band structure points rely on the determination of extended quantum mechanical wavefunctions. But not all important properties of many body systems require explicit knowledge of extended wavefunctions, for example the electronic density and total energy of a system remain unaltered by unitary transformations of the eigenstates, and can be used to calculate the ground state electronic structure of a many body system. The electronic density at a particular point can be evaluated to high accuracy using only information from nearby points in space rather than the knowledge of eigenfunctions extending over the whole of the system. Walter Kohn referred to this locality of properties such as the electronic density as "nearsightedness" [33].

In the next two sections we will describe the properties of Wannier functions and the density matrix, summarising results concerning the locality of these objects in electronic systems.

3.4 Wannier Functions

Wannier functions are localised functions which span the same space as the eigenstates of the electronic band from which they are constructed. They are formed by Fourier transforming the Bloch eigenstates of a system, and are convenient objects for the description of electronic structure [34], [35].

Bloch's theorem tells us that the eigenstates of the Hamiltonian operator $\hat{H}$ will also be eigenstates of the translation operator $\hat{T}$ (i.e. translation by a crystal lattice vector) and that following this the eigenstates can be expressed in the form

$$\psi_i^{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}.\mathbf{r}} u_i^{\mathbf{k}}(\mathbf{r}), \tag{3.3}$$

where $u_i^{\mathbf{k}}(\mathbf{r})$ is a function with the periodicity of the lattice, i is the band index and $\mathbf{k}$ the wavevector. Wannier functions are obtained by Fourier transformation of these Bloch eigenstates,

$$w_i(\mathbf{r} - \mathbf{T}_m) = \frac{\Omega_{cell}}{(2\pi)^3} \int_{BZ} d\mathbf{k} e^{i\mathbf{k}.(\mathbf{r}-\mathbf{T}_m)} u_i^{\mathbf{k}}(\mathbf{r}). \tag{3.4}$$

The lattice point within the unit cell is denoted by $\mathbf{T}_m$.

In [36] Kohn showed that in one dimensional systems containing band gaps it is possible to obtain exponentially decaying Wannier functions in the tight binding regime, following from this result it is also anticipated that the Wannier functions of three dimensional systems will also show exponential decay

3.5 Single Particle Density Matrix

The one particle density matrix $\rho(\mathbf{r}, \mathbf{r}')$ contains all the information about a quantum mechanical system treated using the independent particle formalism (though , expressed in terms of the Fermi function $f_i = 1/(1+exp(\beta(\epsilon_i - \mu))$ where $\beta = 1/kT$,

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_i f_i \psi_i(\mathbf{r}) \psi_i^*(\mathbf{r}'), \tag{3.5}$$

where the sum i is over the number of occupied states (for T greater than 0 this will be more than the total number of electrons). At zero temperature the Fermi distribution function becomes a Heaviside (step) function and then single particle density matrix becomes,

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{N_{el}} \psi_i(\mathbf{r}) \psi_i^*(\mathbf{r}'). \tag{3.6}$$

All of the terms in the DFT expression for the total energy can be expressed in terms of the single particle density matrix rather than KS eigenfunctions [29], as we will see in section 3.8. The reformulation of these terms is important for implementation of the linear scaling algorithms within the CONQUEST computer program. Another important quantity is the band structure energy which is the sum of the independent particle energies. In terms of the Fermi function and the individual eigenvalues it is

$$E_{bs} = \sum_{i=1} \frac{1}{1 + exp(\beta(\epsilon_i - \mu))} \epsilon_i \tag{3.7}$$

with the index i running over all eigenstates. The band structure energy can also be written in terms of the KS Hamiltonian operator,

$$E_{bs} = Tr\left(\frac{1}{1 + exp(\beta(\hat{H} - \mu))}\hat{H}\right), \tag{3.8}$$

this can be equivalently expressed as the trace of the product of the KS Hamiltonian and density matrices since the sum of the eigenvalues of a matrix is its trace,

$$E_{bs} = Tr(\hat{\rho}\hat{H}). \tag{3.9}$$

We can also write down the grand potential in terms of the band energy and the chemical potential,

$$\Omega = E_{bs} - \mu N_{el}, \tag{3.10}$$

which has the property of remaining unchanged if the external potential is altered by an additive constant. Although the associated potential energy will be increased by a term which is proportional to N_{el} when a constant is added, the chemical potential must also increase to conserve the total electron number, and these two changes cancel each other in equation 3.10. This definition of the grand potential will be important later on in section 3.7 when we examine O(N) strategies for minimising the density matrix to obtain the ground state energy.

Cloizeaux established the exponential decay rate of the zero temperature density matrix in insulators [37]; this decay underlies the assumption that truncating the density matrix beyond a certain cut off region is a good approximation as the elements become negligibly small with increasing separation of its arguments, $\mathbf{r}$ and $\mathbf{r}'$, an approximation used for example in CONQUEST to build a sparse Hamiltonian matrix. For metallic systems the

decay rate of the zero temperature density matrix was established by March et al [38] who found it to fall off algebraically. When the metal is analysed at finite temperature the decay rate is found to be exponential [39], [40], which further supports strategies enforcing localisation of the density matrix.

3.6 Building the Hamiltonian

The key to linear scaling DFT calculations lies in constructing sparse Hamiltonian and overlap matrices, so that the number of non zero elements grows in direct proportion to the system size, and then solving the resulting KS equations in a linear scaling manner. The real-space Hamiltonian will be sparse if the basis functions on which it is expressed are cut-off beyond a radius which is less than the system size. A well favoured choice of basis for this purpose is a set of localised orbitals, although Galli and Parrinello [41] have shown how plane waves can also be transformed into a localised form to allow linear scaling. Many different approaches have been adopted towards linear scaling, falling into the two categories of non-variational approaches (which may return an energy below the ground state energy of the system) and variational approaches which will not overestimate the ground state energy (i.e. assign it greater stability). For a review of the various linear scaling strategies we refer the reader to the literature [13] [15]. In section 3.7 we give a description of a strategy based on density matrix minimisation, which is most relevant to CONQUEST [42] [7] [29].

3.7 Variational Density Matrix Approaches

There are two properties that the density matrix must satisfy for a system at absolute zero, one is that it must be a projection operator, and the other is that all eigenvectors with eigenvalue one must correspond to the occupied eigenvectors of the Hamiltonian. To drive an approximate initial density matrix towards the correct form at the ground state a method suggested by Li, Nunes and Vanderbilt [8] may be used, based on a transform known as "Mcweeny purification" [43] which is applied as follows; let ρ be an approximate trial density matrix with eigenvalues in the range $[0, 1]$. Then the quantity $3\rho^2 - 2\rho^3$ will be an improved approximation to the ground state density matrix, having eigenvalues which are driven closer to 1 or 0 depending on whether they correspond to occupied or unoccupied eigenvectors of the Hamiltonian. Below we plot in graph 3.1 the function 3.11

$$y = 3x^2 - 2x^3, \tag{3.11}$$

which demonstrates the maximum and minimum at 1 and 0 respectively, towards which the eigenvalues are driven.

Repeated iteration of the McWeeny transformation upon a trial density matrix will "purify" its form towards that of the actual ground state density matrix. The eigenvalues of the trial density matrix must lie between -1/2 and 3/2 in order for the purification algorithm to drive it towards idempotency, as can be seen from the graph 3.1. Mcweeny purification does not however guarantee that the eigenvectors will correspond to the lowest energy states because imposition of a spatial cut off on the density matrix makes it non variational. The expression for the grand potential at T=0 is used in a

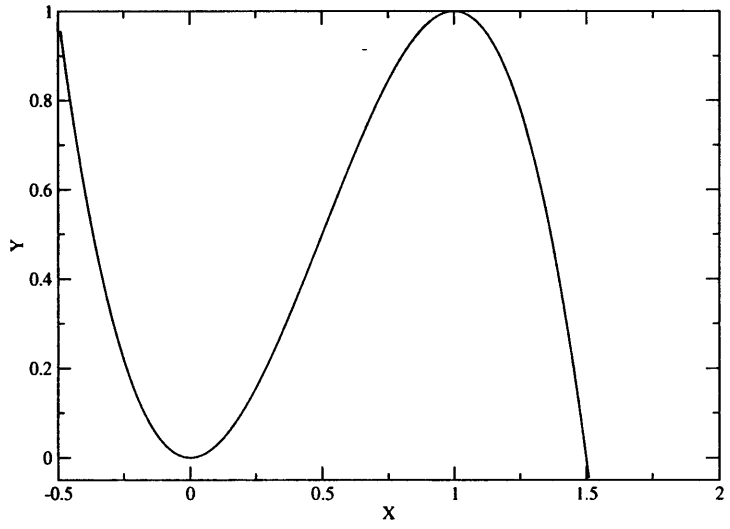

Figure 3.1: Graph showing the function $y = 3x^2 - 2x^3$.

modified form instead;

$$\Omega_s = tr(\rho'(H - \mu)) \approx tr(3\rho^2 - 2\rho^3)(H - \mu). \tag{3.12}$$

We can see that the functional above does not involve local minima because it is a cubic polynomial in all its variables. If we consider two minima in the functional then look along the line which contains them we would again find two minima. However this contradicts the fact that the polynomial is cubic since by definition it cannot have this many minima. One could also worry about the fact that the gradient vanishes independently of the chemical potential when the density matrix has the correct form. But the introduction of fractional occupation numbers prevents this as

$$\rho = \sum_l n_l \langle \psi_l | \psi_l \rangle \tag{3.13}$$

the gradient of the grand potential becomes

$$\frac{\partial \Omega}{\partial \rho} = \sum_l 6(\epsilon_l - \mu) n_l (1 - n_l) \langle \psi_l | \psi_l \rangle. \tag{3.14}$$

This functional is minimised by the correct ground state density matrix, so that the approach is a variational one. A search algorithm using the expression for the gradient of the grand potential in terms of the density matrix can be used for the minimization,

$$\begin{aligned} \frac{\partial \Omega_s}{\partial \rho} &= 3[\rho(H-\mu) + (H-\mu)\rho] \\ &- 2[\rho^2(H-\mu) + \rho(H-\mu)\rho] + (H-\mu)\rho^2] \end{aligned} \tag{3.15}$$

As long as ρ does not have eigenvalues outside the range $[0, 1]$ the algorithm is stable. An expensive requirement of this approach is the multiplication of matrices of dimension N_{basis}, so it is quicker for basis sets comprising only a small number of functions.

3.7.1 Li, Nunes and Vanderbilt method

This method is relevant to the density matrix minimisation scheme within CONQUEST [5], it is essentially the strategy described in the previous section adapted to the tight binding context. The Li, Nunes and Vanderbilt (LNV) scheme [8] provided a variational solution for the density matrix, which was truncated to zero beyond a preselected cut off radius R_C. The solution of the variational problem involved an unconstrained minimization, using a conjugate gradients search algorithm.

LNV considered a unit supercell having N atoms, each with M basis functions

at its centre. The density matrix is defined as

$$\rho_{ij} = \sum_n \psi^*_{ni}\psi_{nj} \tag{3.16}$$

where the n index labels occupied eigenstates of the Hamiltonian, and the i, j label the different basis functions. So the Schrodinger equation looks like

$$\sum_{ij} H_{ij}\psi_{n,j} = \epsilon_n\psi_{ni} \tag{3.17}$$

Because ρ is a projection operator onto the space of occupied states it will satisfy $\rho^2 = \rho$, the idempotency condition. We have already seen that the exponential decay of ρ in insulators and algebraic decay in metals can be used to justify its truncation beyond a preselected radial cut off in section 3.3. Within a periodic system the density matrix will be invariant with respect to translation by a common lattice vector,

$$\rho_{ij} = \rho_{i'j'} \tag{3.18}$$

Thus the unique elements of ρ may be enumerated by allowing i to run over the NM orbitals in a single unit cell, and restricting j to span only those LM orbitals which are within the cut-off radius R_c of the atom centred at i. Thus the number of degrees of freedom in the density matrix becomes $N * L * M^2$ which we see is linear in N, the number of atoms in the super cell.

Before one can proceed with a straightforward minimisation of

$$E = tr(\rho H) = \sum_{ij} \rho_{ij}H_{ji} \tag{3.19}$$

one must also make sure that the idempotency constraint will be enforced, otherwise the eigenvalues corresponding to occupied orbitals will diverge towards $+\infty$ and those eigenvalues corresponding to eigenvectors lying above the chemical potential will diverge towards $-\infty$. However if the eigenvalues of ρ are constrained to lie in the range $[0, 1]$ the minimization procedure will drive them towards either one or zero.

Thus allowing ρ' to be the physical density matrix and ρ to be the trial density matrix LNV mimise the following functional,

$$\Omega = tr[\rho'(H - \mu)] = tr[(3\rho^2 - 2\rho^3)(H - \mu)] \tag{3.20}$$

No constraint is explicitly imposed, instead LNV search for a local minimum of Ω where the eigenvalues cluster around $0, 1$. Usually for purposes of initial input the density matrix is set equal to one half the identity matrix. The gradient of Ω w.r.t. ρ can be written

$$\frac{\partial \Omega}{\partial \rho} = 3(\rho H' - H'\rho) - 2(\rho^2 H' + \rho H' \rho + H'\rho^2) \tag{3.21}$$

where $H' = H - \mu$. The gradient in equation 3.21 above might then (naïvely) be used to perform a conjugate gradients minimisation of the LNV functional to obtain the ground state density matrix and energy. However, it has been discovered [44] that it is in fact tensorially incorrect to use this expression for the gradient to update the density matrix, and a different expression must be used instead, necessary in avoiding the problems which would otherwise occur. The expression for the correct gradient Φ is in fact

$$\Phi_{ij} = (S^{-1}FS^{-1})_{ij} \tag{3.22}$$

where

$$F_{ij} = -\frac{\partial \Omega}{\partial \rho_{ij}}. \tag{3.23}$$

We will see that CONQUEST uses both the LNV density matrix minimisation procedure and direct McWeeny purification of ρ in the search for the ground state for reasons discussed in section 3.9.

3.8 Density Matrix DFT in CONQUEST

The CONQUEST implementation of a linear scaling density matrix minimisation method to find the ground state energy of a many atom system is discussed in [5][7][29][42][45]. Here we show how the KS Hamiltonian in CONQUEST is recast in terms of the density matrix. We showed in the previous chapter that the total energy of a system of nuclei and their orbiting electrons can be expressed as

$$E_{tot} = E_k + E_{ps} + E_H + E_{xc} + E_{II}. \tag{3.24}$$

The respective terms above being the kinetic, pseudopotential, Hartree and exchange-correlation energies of the electrons and the energy of the atomic cores.

We have shown in chapter two that the Hartree and exchange-correlation energies can be written as explicit functionals of the density,

$$E_H = \frac{1}{2}e^2 \int d\mathbf{r}d\mathbf{r}' n(\mathbf{r})n(\mathbf{r}')/|\mathbf{r} - \mathbf{r}'| \tag{3.25}$$

describing the Hartree energy and

$$E_{xc} = \int d\mathbf{r} n(\mathbf{r}) \epsilon_{xc}[n(\mathbf{r})] \tag{3.26}$$

In the conventional formulation of DFT the ground state implies that occupied orbitals are fully occupied, but when computing the ground state energy it is sometimes convenient to relax this condition and allow orbitals to be partially occupied. For example in finite temperature DFT the orbitals may be fractionally populated according to the Fermi-Dirac distribution. When we have fractional occupation of orbitals we write the electronic number density as

$$n(\mathbf{r}) = 2 \sum_i f_i |\psi_i(\mathbf{r})|^2 \tag{3.27}$$

where the f_i are the fractional occupations, giving us altered expressions for the kinetic and pseudopotential energies;

$$E_k = 2 \sum_{i=1}^{N} f_i \langle \psi_i \left| \frac{-\hbar^2}{2m} \nabla^2 \right| \psi_i \rangle \tag{3.28}$$

and

$$E_{ps} = 2 \sum_{i=1}^{N} f_i \langle \psi_i \left| \hat{V}_{ps} \right| \psi_i \rangle. \tag{3.29}$$

In the density matrix formulation of DFT we assume that we find the same ground state whether we minimize the E_{tot} w.r.t. the ψ_i and the f_i (with the f_i being allowed values from [0,1]) or we minimize it w.r.t. fully occupied states. Recalling the definition of the density matrix as given in equation 3.5 we can write the kinetic and pseudopotential energies in terms of this operator;

$$E_k = -\frac{\hbar}{2m} \int d\mathbf{r} [\nabla_r^2 \rho(\mathbf{r}, \mathbf{r}')]_{\mathbf{r}=\mathbf{r}'} \tag{3.30}$$

$$E_{ps} = 2 \int d\mathbf{r} d\mathbf{r}' V_{ps}(\mathbf{r}', \mathbf{r}) \rho(\mathbf{r}, \mathbf{r}') \tag{3.31}$$

3.8.1 Localisation of the Density Matrix

By imposing a spatial cut-off on the elements of the density matrix $\rho(\mathbf{r}, \mathbf{r}')$, it is made sparse, so that it contains an amount of information scaling linearly with system size. However, the neglect of small elements means that we lose some accuracy, and that the ground state energy will be above the true ground state energy. In chapter five we examine the convergence of order N results with respect to the range of the density matrix towards results obtained using diagonalisation using the PAO basis set. As the range of the density matrix cut-off is extended the ground state energy found using minimisation will converge on that found through diagonalisation. The implementation of a spatial cut-off relies on the fact that the magnitude of elements of the density matrix tend to fall off quickly with separation, so that we can make the approximation

$$\rho(\mathbf{r}, \mathbf{r}') = 0, |\mathbf{r} - \mathbf{r}'| > R_c \tag{3.32}$$

The speed with which the elements of the density matrix decay with increasing separation depends on the material being considered. In one dimensional insulators it has been shown that the density matrix falls off exponentially with increasing distance, and this law is assumed to hold some validity in three dimensions as well (see section 3.3 for references).

Because it is difficult to work with the six dimensional function $\rho(\mathbf{r}, \mathbf{r}')$ we

assume that it is separable and therefore can be written as

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) K_{\alpha\beta} \phi_\beta(\mathbf{r}'). \tag{3.33}$$

The $\phi_\alpha(\mathbf{r})$ functions we call support functions, and the matrix $K_{\alpha\beta}$ must be Hermitian in order to ensure hermiticity of the density matrix. In order to be expressible in the form 3.33 the density matrix must obey the restriction of having a finite number of non zero eigenvalues. We choose the support functions to be non zero within a finite region only, known as the support region, to make the density matrix local. The coefficients $K_{\alpha\beta}$ are set to vanish if the separation of the support functions exceeds a prespecified amount. When we perform calculations we use a set of basis functions to represent the support functions, these could be numerical values on a grid or a combination of pseudo atomic orbitals for example, whose implementation we discuss in the following chapter. Whereas a numerical grid basis can be fully converged to represent the space spanned by the density matrix the PAOs might not fully span the space required for its complete representation.

3.8.2 Eigenvalue Range of the Density Matrix

With the assumption that the density matrix is separable we minimize the energy functional w.r.t. the support functions and the K matrix elements. However, the minimisation must be such that the eigenvalues of the density matrix remain in the range [0,1], in CONQUEST we are dealing with basis functions (and support functions) which may be non orthogonal, and rephrasing of the LNV tight binding scheme is necessary in order to take account of this. Instead of working with the eigenvalues directly we can express

ρ so that it is restricted to have eigenvalues within this range by definition, in the way of the LNV scheme. We write ρ in terms of an auxiliary function,

$$\rho = 3\sigma * \sigma - 2\sigma * \sigma * \sigma, \tag{3.34}$$

σ being the auxiliary function, with the asterisk denoting an integration,

$$C(\mathbf{r}, \mathbf{r}') = \int d\mathbf{r}'' A(\mathbf{r}, \mathbf{r}'') B(\mathbf{r}'', \mathbf{r}'). \tag{3.35}$$

for example being $C = A * B$. The definition of ρ given above means that if the eigenvalues of σ lie in the range $[-\frac{1}{2}, \frac{3}{2}]$, then the eigenvalues of ρ will lie in the range [0,1] as desired. The relationship between the auxiliary density matrix σ and the support functions can be written

$$\sigma(\mathbf{r}, \mathbf{r}') = \sum_{\alpha,\beta} \phi_\alpha(\mathbf{r}) L_{\alpha\beta} \phi_\beta(\mathbf{r}'). \tag{3.36}$$

We then see that the relation between the K matrix in 3.33 and the L matrix above is that

$$K = 3LSL - 2LSLSL \tag{3.37}$$

where S is the overlap between support functions,

$$S_{\alpha\beta} = \int dr \phi_\alpha(r) \phi_\beta(r'). \tag{3.38}$$

To ensure linear scaling the support functions $\phi_\alpha(r)$ must be non-zero only within a localised spatial region, also known as a 'support region'. In addition the matrix elements $L_{\alpha,\beta}$ are set to zero if the distance between the atomic centres at α and β exceeds a separate cut-off called the L_{range}. The

calculation of the ground state energy should become exact as these cut-off distances are increased to infinity.

Thus we have two sets of parameters with respect to which we may minimise the energy, the coefficients of the basis functions which are used to represent the support functions, and the elements of the K matrix. Minimization must be performed subject to the constraint of constant electron number, and the calculations will be variational, that is they will give an energy that is above the exact DFT ground state energy of the system.

3.9 Ground State Search

In order to minimise the band structure energy CONQUEST currently implements a two-stage technique [7] following on from work done by LNV [8] and Palser and Manolopoulos (PM) [46] both of which rely on McWeeny purification as the fundamental algorithm.

In their work PM point out that if the initial trial density matrix ρ_{trial} commutes with the Hamiltonian and has eigenvalues within the range (0,1) then the McWeeny purification procedure will be guaranteed to find ρ_{gs} which minimises the KS energy. If ρ_{trial} commutes with the hamiltonian then all the subsequent ρ produced at each purification step will also commute. PM also give a form for ρ_{trial} which will commute with H_{ks}, defined in terms of the chemical potential μ and the upper and lower bounds to the eigenvalue

spectrum of the hamiltonian, H_{max} and H_{min},

$$\begin{aligned} \rho_{trial} &= \frac{1}{2}\zeta(\mu I - H) + \frac{1}{2}I \\ \zeta &= min\left(\frac{1}{H_{max} - \mu}, \frac{1}{\mu - H_{min}}\right). \end{aligned} \tag{3.39}$$

PM show that, using this form for ρ_{trial}, a minimisation procedure based on McWeeny purification will eventually converge to the ground state [46].

Unfortunately the above analysis does not take into account the fact that truncation errors will be present due to enforced zeroing of the elements of ρ beyond a cut off radius. In each step of the McWeeny purification the next ρ is formed using matrix multiplication of the previous density matrix. The effect of the matrix multiplication is to increase the range of ρ_{out} compared to ρ_{in} and thus ρ_{out} must be truncated back to the original cut off after the purification step. The effects of this final truncation disrupt monotonic convergence towards the ground state, thus the PM method does not in fact provide a way to obtain ρ_{gs} once a finite spatial cut off is imposed on the elements of the density matrix.

The LNV technique for obtaining ρ_{gs} is however variational as opposed to PM's method. Thus the approximate density matrices produced after application of a spatial cut off will not lead to an energy below that of the ground state. As the range of the density matrix is increased the energy will converge towards the true ground state energy obtained by diagonalisation as the representation becomes more and more exact. The LNV method exhibits linear convergence towards the variational ground state, slower than the quadratic convergence of algorithms based directly on McWeeny purification.

In order to take advantage of the faster descent of the McWeeny purification scheme as well as the variational convergence of the LNV algorithm CONQUEST currently supports a two-tier density matrix minimisation approach, using McWeeny purification until truncation errors in ρ cause the band energy to increase at which point the LNV algorithm is activated to converge towards ρ_{gs}.

3.10 Support Functions

The Hamiltonian matrix elements within CONQUEST are not explicitly expressed with respect to a specific type of basis function, but are expressed in terms of more general "support functions", $\phi_{i\alpha}$ which are restricted to within a certain radial cut off. These are the same support functions we have mentioned in section 3.8.2, allowing us to write the density matrix in separable form. The first index i refers to the atom on which the support function is centered, the second identifying the support function from the set centered on that atom.

Furthermore different types of basis functions may be selected to construct the $\phi_{i\alpha}$, and currently CONQUEST supports two types of basis function, B-splines (blip) functions and pseudo atomic orbitals (PAOs). Here we discuss the properties of the B-spline basis [11] in more detail, leaving a discussion of the PAO basis to the next chapter which provides details of their implementation.

The B-spline basis consists of piece-wise continuous cubic polynomials with a finite spatial extent. The use of cubic splines ensures that the functions and

its first two derivatives are continuous up to the cut off radius. The splines themselves are expressed upon grid points with a spacing a, and are non zero only within the range $-2a < x < 2a$. The definition of the function inside this range is

$$\begin{aligned} B(x) &= 1 - \frac{3}{2}x^2 + \frac{3}{4}|x|^3, 0 < |x| < 1 \\ &= \frac{1}{4}(2 - |x|^3), 1 < |x| < 2 \\ &= 0, |x| > 2. \end{aligned} \tag{3.40}$$

The support functions enter the DFT equations through the overlap matrix elements, the kinetic energy matrix and also the potential energy. It is shown that the first two contributions can be calculated analytically [11], however the contribution from the potential operator must be approximated through quadrature over a real space grid. Notably the nonlocal pseudopotential contribution to the energy must be evaluated on the real space summation grid (we will see later in chapter three that this is not the case for PAOs). The quadrature over a real space grid scales with the cube of the number of grid points, becoming prohibitilvely expensive for very fine grid spacings. The memory required to store the grid information also grows rapidly as the accuracy is increased. However the strong advantage of the blip functions is that they are a systematically convergable representation of the density matrix and the accuracy of representation can be quantified for direct comparison with plane wave basis sets too.

One possibility for the future of CONQUEST may be to implement a combined basis of blips and PAOs within the same calculations, taking advantage of the properties of both types of basis functions. PAOs allow us to perform relatively quick, accurate calculations on large systems with a relatively small

number of basis functions. Blips on the other hand are expensive to compute with, but can be systematically converged to a high level of accuracy. Thus a sensible strategy for calculation would be to obtain the ground state charge and total energy using PAOs and then to switch over to a blip basis to further improve the accuracy.

3.11 Energies in CONQUEST

We have seen that during an electronic structure calculation using CONQUEST a number of different parameters may be varied in order to reach a variational upper-bound to the true ground state energy and density. The key parameters are the elements of the K matrix and also the coefficients relating support functions to the basis set. A good summary of the hierarchy of total energy approximations available within CONQUEST is presented in [45], and the current status of the code is reported in [5].

The search for the ground state charge density implemented in CONQUEST consists of three different loops, nested within one another. The first (innermost) loop within CONQUEST minimises the K (density) matrix whilst holding the basis function coefficients fixed. The second (middle) loop concerns self-consistency of the charge density used to form the KS potential and that obtained by solving the KS equation. The third (outermost) loop then concerns minimisation of the basis function coefficients. When all three loops are used the resulting forces and energies can be made to converge towards exact diagonalisation results, which we demonstrate with pseudo atomic orbtitals in chapter five.

The local orbital formulation of DFT, as we have seen above, allows for a natural sequence of searches for the ground state energy in which fewer or more sets of parameters can be varied according to the desired accuracy. The implementation of the methods involving local orbital description can also use either exact diagonalization of the Hamiltonian matrix in order to find the ground state KS orbitals or a linear scaling method such as the variation of a localised density matrix as in CONQUEST. Hence in full ab initio DFT minimisation w.r.t. K and the basis set coefficients is done and the KS density is obtained self consistently [45]. In self-consistent ab initio tight binding (SC-AITB) the basis coefficients are held fixed. If self-consistency is also dispensed with the resulting total energy is referred to as the non self-consistent ab initio tight binding (NSC-AITB) energy.

Chapter 4

Pseudo Atomic Orbitals

4.1 Introduction

In this chapter we discuss the development and implementation of pseudo atomic orbitals (PAOs) within the CONQUEST computer code. The popularity and success of the PAO based SIESTA [10] program, reflected in the 462 publications listed on the SIESTA webserver, made a convincing case for their incorporation into CONQUEST where they could be used within a different approach to linear scaling DFT. The B-spline, or blip, basis set already inside CONQUEST has properties which should enable it to be used in a complementary manner to PAOs. For example PAOs could be used in the initial stages of a DFT calculation before the systematically convergable blips are used to improve the estimate of the total energy.

The computer code used to calculate overlap matrix elements between PAO functions within CONQUEST was developed from scratch by the author,

taking as a starting point the reformulation of the overlap integral between two PAOs from real space into reciprocal space as described within ref. [10]. This approach enables the use of fast fourier transforms (FFTs) for quick evaluation of the matrix elements, which are obtained and tabulated before the DFT computation is begun. That they are calculated only at the start and then looked up rather than repeatedly evaluated saves valuable processor time during the electronic structure calculation.

In the next section 4.2 we give a brief introduction to the properties and nomenclature of PAO basis sets before presenting details on how they enter into expressions for the total energy and forces. In section 4.3 details of how the PAOs are created are given, in short the machinery provided by SIESTA is used and more details can be found in [10] [47]. In the remainder of the chapter we discuss the algebra involved in evaluating the overlap integrals 4.4, which are used in expressions for the total energy, and also for the PAO function gradients which are used in obtaining the forces acting on the ions.

4.2 The PAO Basis Set

A set of PAOs $\phi_\alpha(\mathbf{r} - \mathbf{R}_i)$ are typically associated with an atom centred at $\mathbf{R}_i$, and at a given atomic site we can write this set as

$$\phi_\alpha(\mathbf{r}) = \phi_{nlm}(\mathbf{r}) = \phi_{nl}(\mathbf{r})Y_{lm}(\Omega). \tag{4.1}$$

Here the index n denotes the set of radial functions which are associated with an angular momentum channel l, Ω denotes the solid angle. Such an arrangement is called a 'multiple-zeta' basis set. Having one radial function

per l channel gives us a single-zeta basis, two radial functions a double-zeta basis and so on. We can ensure that the basis functions are real (as opposed to spherical harmonics which are complex quantities) by using appropriate linear combinations,

$$\begin{aligned} \chi^{+}_{lm} &= \frac{1}{\sqrt{2}}(Y_{lm} + Y^{*}_{lm}) \\ \chi^{-}_{lm} &= \frac{1}{\sqrt{2}i}(Y_{lm} - Y^{*}_{lm}) \end{aligned} \tag{4.2}$$

Matrix elements of local potentials (Hartree, exchange-correlation) can be treated in the manner of plane-wave calculations where the terms are evaluated on a grid. The wavefunctions in the localised orbital basis must be transferred onto this grid for the summation of the matrix elements, which is a relatively straightforward operation.

The overlap matrix elements,

$$S_{i\alpha j\beta} = \langle \phi_{i\alpha} | \phi_{j\beta} \rangle, \tag{4.3}$$

are necessary in forming the density kernel K, as discussed in the previous chapter;

$$K = 3LSL - 2LSLSL. \tag{4.4}$$

In order to form the Hamiltonian we must also evaluate matrix elements of the kinetic energy operator and the non local pseudopotential operator. The non local part of the pseudopotential operator can be expressed in terms of projector functions as in section 2.11 [45], so that the Hamiltonian matrix

elements appear as

$$H_{\alpha\beta}^{NL} = \sum_{klm} \langle \phi_\alpha | \chi_{klm} \rangle \langle \chi_{klm} | \phi_\beta \rangle. \tag{4.5}$$

Because the Kleinman-Bylander projector functions have an identical form to the PAOs [24] the non local pseudopotential matrix elements can be calculated using the same kernel of code as for the overlap matrix elements. We will also see that the kinetic energy matrix can be easily computed using the same machinery too.

When computing forces we must also concern ourselves with the gradients of PAO functions. For example the total force on an ion (k) in full ab initio DFT can be written [45]

$$F_k = F_k^{PS} + F_k^{Pulay} + F_k^{II}. \tag{4.6}$$

It is the sum of a pseudopotential force, the Pulay force (which arises due to the use of an atom centred basis set) and the ion-ion interaction. The expression for the Pulay force requires calculation of terms like $\langle \nabla_k \phi_\alpha | \phi_\beta \rangle$ and $\langle \phi_\alpha | \nabla_k \phi_\beta \rangle$. We must also evaluate gradients of the support-projector matrix elements to enumerate F_k^{PS}. The machinery for computing such gradients of PAO functions is discussed in section 4.11.

4.3 Constructing PAO Functions

Pseudo atomic orbitals are the eigenfunctions of a pseudo-atom, that is an atom treated using the pseudopotential approximation to describe the core

electrons, confined within a given radius. The method for generating the PAOs is currently identical to that used in the SIESTA code as we used the SIESTA Gen-Basis utility to create our PAOs. We describe the details of their PAO generation method here, essentially following the description provided in the SIESTA review paper [10]. As mentioned before the PAO basis functions take the form of a radial function multiplied by real spherical harmonic combinations, which are indexed by their orbital and azimuthal angular momentum (l, m). When more than a single radial function is associated with each spherical harmonic the basis is called a 'multiple-zeta' basis set. The accuracy of the basis set representation increases with this radial multiplicity, since each additional function gives us another degree of freedom. However in order to converge the quality of the basis set representation we must include PAOs with increasing numbers of nodes, i.e. of higher and higher energies.

A minimal basis set is one in which there is a single radial function associated with each angular momentum channel l. This kind of basis set is useful for describing elements with sp3 type bonding, like Silicon or Germanium, since their chemical and bulk properties are dominated by their tetrahedral bonding behaviour. Indeed many tight binding parameterisations for Si and Ge assume an explicit form for the four sp3 orbitals similar to our minimal basis set. For more accurate calculations a dzp basis (described later) may be used.

In order to the generate orbitals for the minimal basis set the method of Sankey and Niklewski is used. The orbitals are made equal to the eigenfunctions of the pseudo-atom confined within a spherical potential barrier, set equal to infinity. The first node of an eigenfunction gives the cut-off length

at which it is truncated. The position of this node can be tuned through the energy $\epsilon_l + \delta\epsilon_l$, following the expression

$$\hat{H}_{pseudo-atom}\phi_l = (\epsilon_l + \delta\epsilon_l)\phi_l. \tag{4.7}$$

The energy ϵ_l is the energy obtained using the lth PAO without a hard-sphere radial cut off around the pseudo-atom. The increase $\delta\epsilon_l$ is due to the introduction of a cut of radius to the PAO. Instead of fixing an identical cut-off length for all the orbitals one can also set a common energy shift $\delta\epsilon_l$. Truncating the PAOs necessarily raises the obtained ground state energy of the pseudo atom since the degree of freedom in the basis set is reduced. Therefore one would like to know, when using PAOs with a finite radial cut off, the relative error or shift in energy per atom resulting from the cut off R_c, which is why it is also used as a criterion for selecting PAOs.

Having constructed the minimal basis set we can increase its multiplicity by including additional radial functions for each l channel. One way to do this would be to include radial functions having more nodes, which will also have greater energies than the functions in the minimal set. As mentioned such an expansion will be systematically convergent, but will also be computationally expensive, due to the rapidly increasing radial extent of the PAOs. We could lose the locality of the basis set quite quickly, which would slow down our calculations a lot. Instead the method of SIESTA is used, and it is based on a technique found in quantum chemistry, where localised basis sets are more usual, called 'split-valence'. Often in quantum chemical calculations the minimal basis set comprises a fixed pre-chosen linear combination of Gaussians, all of which decay radially with different rates, forming a single radial channel. To form a double zeta orbital a single Gaussian may be 'split'

from the original linear combination (usually selected to be the most slowly decaying Gaussian) to form its own radial channel. The higher zeta orbitals can then be constructed simply by separating out more Gaussians from the original linear combination. However for numerical PAO functions the second zeta orbitals are constructed to share the same tail behaviour as the single zeta functions, but are given a simpler polynomial behaviour inside what is called the 'split-radius', r_l^s.

$$\begin{aligned} \phi_l^{2\zeta}(r) &= r^l(a_l - b_l r^2), r < r_l^s \\ &= \phi_l^{1\zeta}(r), r >= r_l^s. \end{aligned} \tag{4.8}$$

The constants a_l and b_l are chosen to ensure continuity of the function and its first derivative at r_l^s. The smooth behaviour of higher-zeta orbitals within r_l^s helps to ensure speedy evaluation of matrix elements. The split-radius is set by fixing the norm of the minimal zeta functions within it to be roughly 0.15, this is a heuristic setting due to the SIESTA team [47]. In [47] they demonstrate that quantities obtained using a split norm of 0.15 do not differ much from those obtained using an optimized split norm value (i.e. that value which gives the lowest energy - in the paper they find a difference in cohesion energy of bulk Si to be 0.2 eV per pair). They find the optimized split norm values to vary from 0.1 to 0.2, showing no great deviation from the default setting of 0.15. An exception to the above rule occurs for the hydrogen atom, for which a split norm of 0.5 becomes necessary.

As well as the valence orbitals of the confined pseudo-atom it is also necessary to include polarization orbitals in the basis set, in order to allow description of the distortion of atomic orbitals during bond-formation. A polarization orbital is formed by applying an electric field to an atomic orbital thus al-

tering the electronic charge distribution to be more like that of a bonding orbital, where charge accumulates into the bond itself. During bond making and breaking electrons may be found in higher energy states than in the relaxed ground state atomic configuration. The polarization orbital is of a higher energy than the multiple-zeta set, and should provide a more suitable description of the excited electron. Simply including PAOs of higher angular momenta than those already in the basis can result in orbitals which are inconveniently long-ranged as these functions will contain more and more nodes, so instead a polarization orbital is formed from the valence PAO of highest l. A small electric field B is applied along the z-axis resulting in a change in the original orbital,

$$(H - E)\delta\phi = -(\delta H - \delta E)\phi. \tag{4.9}$$

with $\delta H = Bz$. The perturbed orbital will expand out with angular orbital momenta $l' = l \pm 1$ and $m' = m$. Because there will probably already be an orbital with $l - 1$ included in the basis set only the $l + 1$ term of this polarization orbital is retained for inclusion.

4.4 PAO Matrix Elements and Gradients

In the sections which follow we will write down all the relevant expressions for matrix elements between PAO functions, and also PAO gradients, which are now used in CONQUEST. We begin by presenting the overlap integral between two PAOs using complex spherical harmonics (section 4.5), as found in [10]. Next we discuss the properties of the spherical Bessel functions (section 4.6) which are key in evaluating the overlap integral, and how these

were implemented within CONQUEST.

The decision was made in developing CONQUEST to use strictly real (not complex) quantitities where possible so that the PAOs actually used are in terms of real combinations of spherical harmonics with the appropriate angular momenta. Working in terms of these combinations alters the final expressions for the overlap integral and these are evaluated explicitly for different cases of angular momenta.

Having developed all the necessary expressions required for PAO matrix elements (section 4.9) we then discuss the calculation of PAO gradients which are important for force evaluation (section 4.11). The code for evaluating both matrix elements and gradients is then tested using PAOs with Gaussian radial functions, whose overlap integrals can be evaluated analytically in order to test the numerical machinery. Finally a consistency test between the total energy and force within CONQUEST is used to demonstrate the accurate implementation of the new basis set within CONQUEST.

4.5 Overlap Integral

The overlap integral between two functions is defined to be

$$S(\mathbf{R}) = \int \psi_1^*(\mathbf{r})\psi_2(\mathbf{r}-\mathbf{R})\mathbf{d}^3\mathbf{r}. \tag{4.10}$$

We are concerned with calculating the overlap integral between PAO functions, radial functions multiplied by real combinations of spherical harmonics, but at this stage we will consider the overlap integral for complex PAOs which

are the product of a real radial part and a complex spherical harmonic. This calculation is discussed in [10] where its implementation in SIESTA is also discussed.

$$\psi_{lm}(\mathbf{r}) = f(r)Y_{lm}(\hat{\mathbf{r}}) \tag{4.11}$$

Using the definition of the Fourier transformation (in three dimensions)

$$\psi(\mathbf{k}) = \frac{1}{(2\pi)^{\frac{3}{2}}} \int \psi(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}} \mathbf{d}^3\mathbf{r} \tag{4.12}$$

$$\psi(\mathbf{r}) = \frac{1}{(2\pi)^{\frac{3}{2}}} \int \psi(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{r}} \mathbf{d}^3\mathbf{k} \tag{4.13}$$

we can re-express the overlap integral in Fourier space, where it becomes diagonal, in accordance with the convolution theorem, which states that a convolution of two functions in real space becomes a product of the same functions when transformed into Fourier space. Inserting the Fourier transform of the second function we have

$$\begin{aligned} S(\mathbf{R}) &= \int \psi_1^*(\mathbf{r}) \left(\frac{1}{(2\pi)^{\frac{3}{2}}} \int \psi_2(\mathbf{k}) e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})} \mathbf{d}^3\mathbf{k} \right) \mathbf{d}^3\mathbf{r} \\ &= \int \int \left(\frac{1}{(2\pi)^{\frac{3}{2}}} \psi_1^*(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}} \mathbf{d}^3\mathbf{r} \right) \psi_2(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{R}} \mathbf{d}^3\mathbf{k} \end{aligned} \tag{4.14}$$

But we recall that

$$\psi^*(\mathbf{k}) = \frac{1}{(2\pi)^{\frac{3}{2}}} \int \psi^*(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}} \mathbf{d}^3\mathbf{r} \tag{4.15}$$

so that

$$S(\mathbf{R}) = \int \mathbf{d}^3\mathbf{k} \psi_1^*(\mathbf{k}) \psi_2(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{R}} \mathbf{d}^3\mathbf{k}. \tag{4.16}$$

Now, we may take the above expression further by introducing the expansion of the plane-wave in terms of spherical harmonics multiplied by spherical Bessel functions (which we will look at in more detail later on);

$$e^{i\mathbf{k}\cdot\mathbf{r}} = \sum_{l=0}^{\infty}\sum_{m=-l}^{m=+l} 4\pi(i)^l j_l(kr)Y_{lm}^*(\hat{\mathbf{k}})Y_{lm}(\hat{\mathbf{r}}), \tag{4.17}$$

here the $j_l(kr)$ are the spherical Bessel functions, which we shall discuss in section 4.6. Inserting this expansion into our Fourier transformed PAOs;

$$\begin{aligned}\psi_1^*(\mathbf{k}) &= \frac{1}{(2\pi)^{\frac{3}{2}}}\int \psi_1^*(\mathbf{r})\left(\sum_{l=0}^{\infty}\sum_{m=-l}^{m=+l} 4\pi(i)^l j_l(kr)Y_{lm}^*(\hat{\mathbf{k}})Y_{lm}(\hat{\mathbf{r}})\right)\mathbf{d^3r} \\ &= \frac{1}{(2\pi)^{\frac{3}{2}}}\int f_1(r)Y_{l_1m_1}^*(\hat{\mathbf{r}}) \\ &* \left(\sum_{l=0}^{\infty}\sum_{m=-l}^{m=+l} 4\pi(i)^l j_l(kr)Y_{lm}^*(\hat{\mathbf{k}})Y_{lm}(\hat{\mathbf{r}})\right)\mathbf{d^3r}.\end{aligned} \tag{4.18}$$

In the second line above we can use the fact that the PAO radial tables are real. The spherical harmonics are all orthonornal to each other;

$$\int d\Omega Y_{l_1m_1}^*(\hat{\mathbf{r}})Y_{l_2m_2}(\hat{\mathbf{r}}) = \delta_{l_1l_2,m_1m_2}. \tag{4.19}$$

so that the infinite summation in the expression 4.18 is reduced to

$$\psi_1^*(\mathbf{k}) = \frac{1}{(2\pi)^{\frac{3}{2}}}\int f_1(r)\left(4\pi(i)^{l_1}j_{l_1}(kr)Y_{l_1m_1}^*(\hat{\mathbf{k}})\right)r^2dr. \tag{4.20}$$

Now (l_1m_1) and (l_2m_2) are the angular momentum indices corresponding to PAOs 1 and 2 respectively. By similar means we can show that

$$\psi_2(\mathbf{k}) = \frac{1}{(2\pi)^{\frac{3}{2}}}\int f_2(r)j_{l_2}(kr)r^2dr4\pi((i)^{l_2})^*Y_{l_2m_2}(\hat{\mathbf{k}}). \tag{4.21}$$

Now we reinsert these expressions for the Fourier transforms of the PAO functions back into the expression 4.16 for the overlap integral.

$$\begin{aligned} S(\mathbf{R}) &= \int \mathbf{d^3k} e^{-i\mathbf{k}.\mathbf{R}} \times \\ &\left(\frac{1}{(2\pi)^{\frac{3}{2}}} \int f_1(r) \left(4\pi(i)^{l_1} j_{l_1}(kr) Y^*_{l_1 m_1}(\hat{\mathbf{k}})\right) r^2 dr\right) \\ &\left(\frac{1}{(2\pi)^{\frac{3}{2}}} \int f_2(r) j_{l_2}(kr) r^2 dr 4\pi((i)^{l_2})^* Y_{l_2 m_2}(\hat{\mathbf{k}})\right) \end{aligned} \tag{4.22}$$

But we can further substitute the expansion of the plane-wave in spherical harmonics into this expression,

$$\begin{aligned} S(\mathbf{R}) &= \int \mathbf{d^3k} \left(\sum_{l=0}^{\infty} \sum_{m=-l}^{m=+l} 4\pi(i)^l j_l(kr) Y^*_{lm}(\hat{\mathbf{k}}) Y_{lm}(\hat{\mathbf{r}})\right) \\ &\left(\frac{1}{(2\pi)^{\frac{3}{2}}} \int f_1(r) \left(4\pi(i)^{l_1} j_{l_1}(kr) Y^*_{l_1 m_1}(\hat{\mathbf{k}})\right)\right) r^2 dr \\ &\left(\frac{1}{(2\pi)^{\frac{3}{2}}} \int f_2(r) j_{l_2}(kr) r^2 dr 4\pi((i)^{l_2})^* Y_{l_2 m_2}(\hat{\mathbf{k}})\right) \end{aligned} \tag{4.23}$$

Equation 4.23 above can again be re-arranged into a more transparent form,

$$\begin{aligned} S(\mathbf{R}) &= \sum_{l=0}^{\infty} \sum_{m=-l}^{m=+l} \int_{all\Omega} d\Omega_k \left(Y_{lm}(\hat{\mathbf{k}}) Y^*_{l_1 m_1}(\hat{\mathbf{k}}) Y_{l_2 m_2}(\hat{\mathbf{k}})\right) \\ &\times \int k^2 dk j_l(kR) \left(\int f_1(r) j_{l_1}(kr) r^2 dr\right) \left(\int f_2(r) j_{l_2}(kr) r^2 dr\right) \\ &\times \left(4\pi(i^l)^* \frac{2}{\pi} (i)^{l_1} ((i)^{l_2})^*\right) \times Y^*_{lm}(\hat{\mathbf{R}}). \end{aligned} \tag{4.24}$$

We know that the integral of the triple spherical harmonics over the solid angle exists only for $|l_1 - l_2| \geq l \leq l_1 + l_2$, so that the sum from l = 0 to ∞ is restricted (the orbital angular momentum quantum numbers obey what is

known as a vector triangle condition), there is also the further condition that $l_1 + l_2 + l$ must equal an even integer.

$$S(\mathbf{R}) = \sum_{l=l_1-l_2}^{l_1+l_2} \sum_{m=-l}^{m=+l} \int d\Omega_k \left((-1)^{m_1} Y_{l_1-m_1}(\hat{\mathbf{k}}) Y_{l_2 m_2}(\hat{\mathbf{k}}) Y_{lm}(\hat{\mathbf{k}}) \right) \left(\int k^2 dk \psi_1'(k) \psi_2'(k) j_l(kR) \right) \left(8(i^l)^*(i^{l_1})(i^{l_2})^* \right) Y_{lm}^*(\hat{\mathbf{R}}). \tag{4.25}$$

In the equation above we have used the shorthand notation

$$\psi_1'(k) = \int r^2 dr j_{l_1}(kR) f_1(r). \tag{4.26}$$

Having obtained this expression for the overlap integral it is then easy to obtain the kinetic energy integrals too, since the real space operator ∇^2 is simply equivalent to a prefactor of k^2 in Fourier space. Thus the kinetic energy integrals are related to the expression for the overlap integral through by a factor of k^2.

4.6 Spherical Bessel functions

When evaluating the overlap integral it is necessary to calculate spherical Bessel transforms of the radial functions. To do this we need to understand better the properties of Bessel functions, particularly their expansion into trigonometric series which allows us to use FFT's instead of integrating the transforms using much slower numerical quadrature. Bessel functions are

solutions to the radial part of the Helmholtz equation, which is

$$\nabla^2\psi - k^2\psi = 0 \tag{4.27}$$

and when we separate it out into spherical polar coordinates we have a radial equation

$$r^2\frac{d^2R}{dr^2} + 2r\frac{dR}{dr} + [k^2r^2 - (n+\frac{1}{2})^2]R = 0 \tag{4.28}$$

the k is the constant from the Helmholtz equation whereas the factor $n(n+1)$ is a separation constant. If we make the substitution $R(kr) = Z(kr)/\sqrt{kr}$ then we regain Bessel's equation

$$r^2\frac{d^2Z}{dr^2} + r\frac{dZ}{dr} + [k^2r^2 - (n+\frac{1}{2})^2]Z = 0 \tag{4.29}$$

but in terms of Bessel functions that are of order $n+\frac{1}{2}$ where n is an integer. These are the spherical Bessel functions. Spherical Bessel functions obey recurrence relations analogous to those of Bessel functions of the first kind, which means that we can compute higher orders of spherical Bessel functions using lower order functions.

$$j_{n-1}(x) + j_{n+1}(x) = \frac{2n}{x}j_n(x) \tag{4.30}$$

However there is another formula, Rayleigh's formula, which allows us to generate the *nth* order spherical Bessel function through repeated differentiation of $J_0(x)$.

$$j_n(x) = (-1)^n\left(\frac{1}{x}\frac{d}{dx}\right)^n(\frac{\sin(x)}{x}) \tag{4.31}$$

Thus we can write down analytically the spherical Bessel functions up to *nth* order using this formula. For the purposes of calculating overlap integrals

between PAO functions we need only evaluate spherical Bessel functions up to order twice that of our PAO of highest l. Having angular momenta up to $l = 3$ (f functions) Bessel functions of at most sixth order will be sufficient. We see using Rayleigh's formula that the zeroth order spherical Bessel function is

$$j_0(x) = \frac{\sin(x)}{x}. \tag{4.32}$$

Because the spherical Bessel functions can be expressed as linear combinations of sine and cosine functions, we can evaluate all integrals of radial functions against them using sums of FFTs against the individual sines and cosines. The explicit forms of the spherical Bessel functions are

$$\begin{aligned}
j_0(x) &= \frac{\sin(x)}{x} \\
j_1(x) &= -\frac{\sin(x)}{x^2} + \frac{\cos(x)}{x} \\
j_2(x) &= 3\frac{\sin(x)}{x^3} - 3\frac{\cos(x)}{x^2} - \frac{\sin(x)}{x} \\
j_3(x) &= 15\frac{\sin(x)}{x^4} - 15\frac{\cos(x)}{x^3} - 6\frac{\sin(x)}{x^2} \\
&+ \frac{\cos(x)}{x} \\
j_4(x) &= 105\frac{\sin(x)}{x^5} - 105\frac{\cos(x)}{x^4} - 45\frac{\sin(x)}{x^3} \\
&+ 10\frac{\cos(x)}{x^2} + \frac{\sin(x)}{x}. \\
j_5(x) &= 945\frac{\sin(x)}{x^6} - 945\frac{\cos(x)}{x^5} - 420\frac{\sin(x)}{x^4} \\
&+ 105\frac{\cos(x)}{x^3} + 15\frac{\sin(x)}{x^2} - \frac{\cos(x)}{x} \\
j_6(x) &= 10395\frac{\sin(x)}{x^7} - 10395\frac{\cos(x)}{x^6} - 4725\frac{\sin(x)}{x^5} \\
&+ 1260\frac{\cos(x)}{x^4} + 210\frac{\sin(x)}{x^3} - 21\frac{\cos(x)}{x^2} - \frac{\sin(x)}{x}.
\end{aligned} \tag{4.33}$$

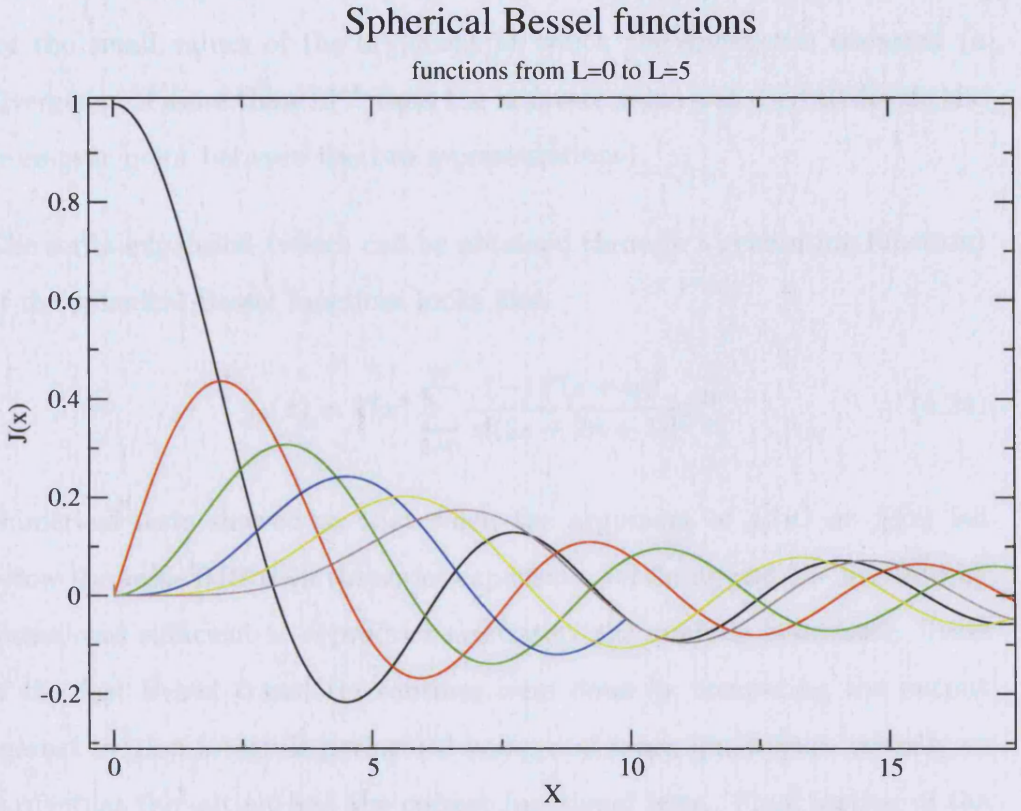

Figure 4.1: We plot above the spherical Bessel functions from $l = 0$ to $l = 5$ produced using the expressions from equations 4.33.

The spherical Bessel functions approach zero as they near the origin, as we can see from the plots for $l = 1$ to $l = 5$, though the $l = 0$ function is unity at the origin (the higher the order of the function the lower its peak magnitude) in figure 4.1. However looking at the equations above we see that as the order of the Bessel function increases we are relying on the exact cancellation of large terms (for example the first two terms of $j_6(x)$ in order to reach zero. Unfortunately double point numerical precision does not provide sufficient accuracy for this to happen if the formulae above are used directly for J_5 and J_6, and we observed a divergence of these functions away zero when they were

computed close to the origin. In order to cure this divergent behaviour we resorted to using the series representation of the spherical Bessel functions for the small values of the argument at which the divergence occurred (a divergence of more than 10^{-6} from the accurate value was used to decide the cross-over point between the two representations).

The series expansion (which can be obtained through a generating function) of the spherical Bessel functions looks like

$$j_n(x) = 2^n x^n \sum_{s=0}^{\infty} \frac{(-1)^s (s+n)!}{s!(2s+2n+1)!} x^{2s}. \tag{4.34}$$

Numerical tests showed us that when the argument of $j_5(x)$ or $j_6(x)$ fell below the value 0.02 then the series expansion, retaining just the five leading terms, was sufficient to reproduce accurately the analytic behaviour. Tests of the fast Bessel transform routines were done by comparing the output against overlap integrals performed using real-space quadrature initially to verify that the output had the correct functional form. Final testing of the overlap integral calculation was done by using Gaussian PAO functions, for which we could calculate the overlap integrals as an analytic function of the vector separation of the two PAOs.

4.7 Overlap Integrals Between Real PAOs

4.7.1 Definition of Real PAOs

The PAO's which are used in CONQUEST are defined to be real, and are set as follows,

$$\psi(\mathbf{r}) = f(r) * C_{lm} P_m^l(\cos(\theta)) \cos(m\phi) \tag{4.35}$$

for m greater than or equal to zero.

$$\psi(\mathbf{r}) = f(r) * C_{lm} P_m^l(\cos(\theta)) \sin(m\phi) \tag{4.36}$$

for m less than zero. Here C_{lm} are normalization constants which ensure that the integral of the square of the PAO angular component over the entire solid angle is unity. The C_{lm} has a contribution from the integral over ϕ and also from the integral over θ. We calculate the normalization constant for the integral over ϕ for m greater than zero as follows;

$$\frac{1}{C_{lm\phi}} = \int_0^{2\pi} \cos^2(m\phi) d\phi \tag{4.37}$$

We evaluate the integral and find that it is equal to π, so that $C_{lm\phi} = 1/\sqrt{\pi}$, for m > 0. Also we know that the normalization factor for the associated Legendre polynomial is

$$C_{lm\theta} = \sqrt{\frac{2l+1(l-m)!}{2(l+m)!}} \tag{4.38}$$

this means the overall normalization constant C_{lm} will be

$$\begin{aligned} C_{lm} &= C_{lm\phi} * C_{lm\theta} \\ &= \sqrt{\frac{2l+1(l-m)!}{2\pi(l+m)!}} \end{aligned} \tag{4.39}$$

To calculate the normalization coefficient for the associated Legendre polynomial we start from the equation

$$\int_{-1}^{1} P_{l_1}^m(x) P_{l_2}^m(x) dx = \frac{2(m+l_1)!}{(2l_1+1)(m-l_1)!} \delta_{l_1 l_2} \tag{4.40}$$

and a change of variable $x = \cos(\theta)$ gives us

$$\int_0^{\pi} P_{l_1}^m(cos(\theta)) P_{l_2}^m(cos(\theta)) sin(\theta) d\theta = \frac{2(m+l_1)!}{(2l_1+1)(m-l_1)!} \delta_{l_1 l_2} \tag{4.41}$$

which is what we require. Above we treated the m greater than zero case only, and did not include the m equal to zero case despite the similar functional form of the PAOs because the integral over $d\phi$ yields 2π rather than π and so the normalization constant becomes

$$C_{l0} = \sqrt{\frac{2l+1(l-m)!}{4\pi(l+m)!}} \tag{4.42}$$

instead. Similarly the normalization constant for the case where m is less than zero, C_{lm-} is equal to

$$C_{lm-} = \sqrt{\frac{2l+1(l-m)!}{2\pi(l+m)!}} \tag{4.43}$$

Now that the normalization constants have been established it is straightforward to show that to construct the real combinations of spherical harmonics

used in the definitions of real PAOs in 4.35 we use complex spherical harmonics as follows,

$$Y_{lm}(\theta,\phi) = B_{lm}P_m^l(\cos(\theta))e^{im\phi} \tag{4.44}$$

where

$$B_{lm} = C_{l0} = \sqrt{\frac{2l+1(l-m)!}{4\pi(l+m)!}} \tag{4.45}$$

so that

$$C_{lm+}P_{m+}^l(cos(\theta)cos(m\phi)) = \frac{1}{\sqrt{2}}(Y_{lm}(\theta,\phi) + Y_{lm}^*(\theta,\phi)) \tag{4.46}$$

for m greater than zero,

$$C_{l0}P_0^l(cos(\theta)cos(m\phi)) = \frac{1}{2}(Y_{l0}(\theta,\phi) + Y_{l0}^*(\theta,\phi)) \tag{4.47}$$

for m equal to zero,

$$C_{lm-}P_{m-}^l(cos(\theta)cos(m\phi)) = \frac{1}{\sqrt{2}}(Y_{l0}(\theta,\phi) - Y_{l0}^*(\theta,\phi)) \tag{4.48}$$

for m less than zero.

4.8 Important Identities

Important identities that we are going to use in the course of our calculations using real PAOs are

$$P_{-m}^l(x) = (-1)^m\frac{(l-m)!}{(l+m)!}P_m^l(x) \tag{4.49}$$

$$Y^*_{lm}(\hat{\mathbf{r}}) = (-1)^m Y_{l-m}(\hat{\mathbf{r}}) \tag{4.50}$$

$$Y_{lm}(\hat{\mathbf{r}}) = \left(\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}\right) P^l_m(cos\theta)e^{im\phi} \tag{4.51}$$

4.9 Matrix Elements Between Real PAOs

There are three distinct cases which we must consider, depending on the sign of the magnetic quantum number m of each of the two PAOs, these are

- m_1 greater than/equal to zero and m_2 less than zero
- both m_1 and m_2 greater than/equal to zero
- both m_1 and m_2 are less than or equal to zero

the difference between these three cases lies in the signs of the complex spherical harmonics which are combined to form the real combinations. We will now calculate the first combination on the list above, with m1 greater than or equal to zero and m2 being less than zero.

4.9.1 m_1 greater than/equal to zero and m_2 less than zero

The overlap integral, according to what we have already worked out, will have the following form;

$$\begin{aligned} S(\mathbf{R}) &= C_{lm+0}C_{lm-}\int f_1(r)\left(Y_{l_1m_1}(\hat{\mathbf{r}})+Y^*_{l_1m_11}(\hat{\mathbf{r}})\right)^* \\ &* \; f_2(r')\left(Y_{l_2m_2}(\hat{\mathbf{r}'})-Y^*_{l_2m_2}(\hat{\mathbf{r}'})\right)\mathbf{d}^3\mathbf{r}, \end{aligned} \tag{4.52}$$

$\mathbf{r}'$ denotes $\mathbf{r}-\mathbf{R}$, with $\mathbf{R}$ the vector separation between the two PAOs. But we know that for complex spherical harmonics the overlap integral comes out as

$$\begin{aligned} S(\mathbf{R}) &= \int f_1(r)Y^*_{l_1m_1}(\hat{\mathbf{r}})f_2(r')Y_{l_2m_2}(\mathbf{r}')\mathbf{d}^3\mathbf{r} \\ &= 8\sum_{l=|l_1-l_2|,2}^{l_1+l_2}(i^{l_1}(i^{l_2*})(i^{l*})\left(\int k^2dkf_1(k)f_2(k)j_l(kR)\right) \\ &\times \int d\omega(-1)^{m_1}Y_{l_1-m_1}(\hat{\mathbf{k}})Y_{l_2m_2}(\hat{\mathbf{k}})Y_{lm}(\hat{\mathbf{k}})(-1)^mY_{l-m}(\hat{\mathbf{R}}). \end{aligned} \tag{4.53}$$

This is a result that we have already established previously by means of Fourier transforms, in the above equation we have $f_1(k)=\int r^2f_1(r)j_l(kr)dr$. Returning to our original problem, 4.52 we notice that there will be a radial integration similar to that of 4.53 but that it will have a more complicated angular component due to the different spherical harmonic products we will get after multiplying two real PAO functions together. Looking at the angular part we find;

$$I_{\Omega_r}=(Y_{l_1m_1}(\mathbf{r})+Y^*_{l_1m_1}(\mathbf{r}))(Y_{l2m_2}-Y^*_{l_2m_2}(\mathbf{r})) \tag{4.54}$$

and the different products of these spherical harmonic functions will map out to give us the following overlap integral angular terms;

$$Y_{l_1m_1}(\mathbf{r})Y_{l_2m_2}(\mathbf{r}) \rightarrow Y_{l_1m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_a}(\mathbf{k})Y^*_{lm_a}(\mathbf{R}). \tag{4.55}$$

$$Y^*_{l_1m_1}(\mathbf{r})Y_{l_2m_2}(\mathbf{r}) \rightarrow Y^*_{l_1m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_b}(\mathbf{k})Y^*_{lm_b}(\mathbf{R}). \tag{4.56}$$

$$Y_{l_1m_1}(\mathbf{r})(-1)Y^*_{l_2m_2}(\mathbf{r}) \rightarrow Y_{l_1m_1}(\mathbf{k})(-1)Y^*_{l_2m_2}(\mathbf{k})Y_{lm_c}(\mathbf{k})Y^*_{lm_c}(\mathbf{R}). \tag{4.57}$$

$$Y^*_{l_1m_1}(\mathbf{r})(-1)Y^*_{l_2m_2}(\mathbf{r}) \rightarrow Y^*_{l_1m_1}(\mathbf{k})(-1)Y^*_{l_2m_2}(\mathbf{k})Y_{lm_d}(\mathbf{k})Y^*_{lm_d}(\mathbf{R}). \tag{4.58}$$

We can show that the terms above are correct using the overlap integral result for complex spherical harmonics that we derived earlier since we know that 4.53 is true we can then use 4.50 to reformulate other spherical harmonic products

$$Y_{l_1m_1}(\mathbf{r})Y_{l_2m_2}(\mathbf{r}) \rightarrow (-1)^{m_1}Y^*_{l_1-m_1}(\mathbf{r})Y_{l_2m_2}(\mathbf{r}). \tag{4.59}$$

Then we have the complex conjugate of the spherical harmonic involving m_1 only and we can apply 4.53 directly to get the corresponding angular term in the overlap integral. In the equations above when the functions of $\mathbf{k}$ are integrated over the solid angle in k space they disappear unless $\delta(m_1 + m_2 + m_3)$ is equal to zero, where m_3 is one of $m_a \rightarrow m_d$. This means that

$$m_a = -m_d \tag{4.60}$$

$$m_b = -m_c \tag{4.61}$$

Now we pair the different angular components together so that we can use the relationship

$$\int d\Omega_k Y_{l_l m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{l_3 m_3}(\mathbf{k}) = \int d\Omega_k Y_{l_l - m_1}(\mathbf{k}) Y_{l_2 - m_2}(\mathbf{k}) Y_{l_3 - m_3}(\mathbf{k}) \tag{4.62}$$

We will pair off the term involving m_a and its counterpart m_d and simplify them;

$$\begin{aligned} I_1 &= Y_{l_1 m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{l m_a}(\mathbf{k}) Y^*_{l m_a}(\mathbf{k}) \\ &+ Y^*_{l_1 m_1}(\mathbf{k}) (-1) Y^*_{l_2 m_2}(\mathbf{k}) Y_{l m_d}(\mathbf{k}) Y^*_{l m_d}(\mathbf{k}) \end{aligned} \tag{4.63}$$

Now we use the relation 4.50 to rewrite the above as

$$\begin{aligned} I_1 &= Y_{l_1 m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{l m_a}(\mathbf{k}) Y^*_{l m_a}(\mathbf{R}) \\ &+ (-1)^{m_1+m_2} Y_{l_1 - m_1}(\mathbf{k}) (-1) Y_{l_2 - m_2}(\mathbf{k}) Y_{l - m_a}(\mathbf{k}) Y^*_{l - m_a}(\mathbf{R}) \\ &= Y_{l_1 m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{l m_a}(\mathbf{k}) \\ &* \left(Y^*_{l m_a}(\mathbf{R}) + (-1)(-1)^{m_1+m_2}(-1)^{m_a} Y_{l m_a}(\mathbf{R}) \right) \\ &+ (-1)^{m_1+m_2} Y_{l_1 - m_1}(\mathbf{k}) (-1) Y_{l_2 - m_2}(\mathbf{k}) Y_{l - m_a}(\mathbf{k}) Y^*_{l - m_a}(\mathbf{R}) \\ &* \left(Y^*_{l m_a}(\mathbf{R}) - Y_{l m_a}(\mathbf{R}) \right). \end{aligned} \tag{4.64}$$

But we see that the combination $Y^*_{l m_a}(\mathbf{R}) - Y_{l m_a}(\mathbf{R})$ is a real spherical harmonic

$$Y^*_{l m_a}(\mathbf{R}) - Y_{l m_a}(\mathbf{R}) = -Pref(l, m_a) P_{l m_a}(cos(\theta)) sin(m_a \phi) \tag{4.65}$$

where $Pref(l, m_a)$ is the appropriate normalisation prefactor. We can also treat the combination of angular components involving $m_1 - m_2$ in a similar

way to find

$$\begin{aligned} I_2 &= (-1)^{m_1} Y_{l_1 -m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{l m_b}(\mathbf{k}) Y^*_{l m_b}(\mathbf{R}) \\ &+ (-1) Y_{l_1 m_1}(\mathbf{k}) (-1)^{m_2} Y_{l_2 -m_2}(\mathbf{k}) Y_{l -m_b}(\mathbf{k}) Y^*_{l -m_b}(\mathbf{R}) \end{aligned} \tag{4.66}$$

and we then notice that

$$(-1)^{m_1} Y^*_{l m_b}(\mathbf{r}) + (-1)^{1+m_2} (-1)^{m_b} Y_{l m_b}(\mathbf{R}) = (-1)^{m_1} \left(Y^*_{l m_b}(\mathbf{R}) - Y_{l m_b}(\mathbf{R}) \right) \tag{4.67}$$

which gives rise to a real spherical harmonic as we saw in 4.65. So the overlap integral between these two PAOs will involve a radial integration identical to that of 4.53 but multiplied by an angular factor which is the sum of the two components given above;

$$\begin{aligned} S(\mathbf{R}) &= 8 \sum_{l=|l_1 - l_2|,2}^{l_1 + l_2} (i^{l_1} (i^{l_2 *}) (i^{l *}) \left(\int k^2 dk f_1(k) f_2(k) j_l(kR) \right) \\ &\times \int d\Omega_k (Y_{l_1 m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{l m_a}(\mathbf{k}) \\ &\times (-pref(l, m_a) P_{l m_a}(cos(\theta)) sin(m_a \phi)) \\ &+ (-1)^{m_1} y_{l_1 -m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{l m_b}(\mathbf{k}) \\ &\times (pref(l, m_b) P_{l m_b}(cos(\theta)) sin(-m_b \phi))). \end{aligned} \tag{4.68}$$

For the overlap integrals involving PAO combinations where both m_1 and m_2 are either less than zero or greater than/equal to zero only the angular part of the overlap integral will differ from that for the case presented above; and we can evaluate these angular parts using the same technique.

4.9.2 Both m1 and m2 greater than or equal to zero

Now for the overlap integral we are looking at a combination of spherical harmonics that looks like

$$(Y_{l_1m_1}(\mathbf{r}) + Y^*_{l_1m_1}(\mathbf{r})) * (Y_{l_2m_2}(\mathbf{r}) + Y^*_{l_2m_2}(\mathbf{r})). \tag{4.69}$$

Now the angular contributions to the overlap integral will look like;

$$Y_{l_1m_1}(\mathbf{r})Y_{l_2m_2}(\mathbf{r}) \rightarrow Y_{l_1m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_a}(\mathbf{k})Y^*_{lm_a}(\mathbf{R}). \tag{4.70}$$

$$Y^*_{l_1m_1}(\mathbf{r})Y_{l_2m_2}(\mathbf{r}) \rightarrow Y^*_{l_1m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_b}(\mathbf{k})Y^*_{lm_b}(\mathbf{R}). \tag{4.71}$$

$$Y_{l_1m_1}(\mathbf{r})Y^*_{l_2m_2}(\mathbf{r}) \rightarrow Y_{l_1m_1}(\mathbf{k})Y^*_{l_2m_2}(\mathbf{k})Y_{lm_c}(\mathbf{k})Y^*_{lm_c}(\mathbf{R}). \tag{4.72}$$

$$Y^*_{l_1m_1}(\mathbf{r})Y^*_{l_2m_2}(\mathbf{r}) \rightarrow Y^*_{l_1m_1}(\mathbf{k})Y^*_{l_2m_2}(\mathbf{k})Y_{lm_d}(\mathbf{k})Y^*_{lm_d}(\mathbf{R}). \tag{4.73}$$

Now when we pair together the angular terms we will get

$$\begin{aligned} I_1 &= Y_{l_1m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_a}(\mathbf{k})Y^*_{lm_a}(\mathbf{R}) \\ &+ Y^*_{l_1m_1}(\mathbf{k})Y^*_{l_2m_2}(\mathbf{k})Y_{l-m_a}(\mathbf{k})Y^*_{l-m_a}(\mathbf{k}) \end{aligned} \tag{4.74}$$

we see that these will combine as

$$\begin{aligned} I_1 &= Y_{l_1m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_a}(\mathbf{k}) \\ &\times \left(Y^*_{lm_a}(\mathbf{R}) + Y_{lm_a}(\mathbf{R})\right). \end{aligned} \tag{4.75}$$

The bracketed quantity is again just a real spherical harmonic which looks like

$$Y^*_{lm_a}(\mathbf{R}) + Y_{lm_a}(\mathbf{R}) = pref(l, m_a) P_{lm_a}(cos(\theta)) cos(m_a \phi). \tag{4.76}$$

The second angular component has the form

$$\begin{aligned} I_2 &= Y_{l_1 -m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{lm_b}(\mathbf{k}) \\ &\times \left((-1)^{m_1} Y^*_{lm_b}(\mathbf{R}) + (-1)^{m_2} Y^*_{l-m_b}(\mathbf{R}) \right). \end{aligned} \tag{4.77}$$

The bracketed quantity again simplifies into a real spherical harmonic since $m_b = m_1 - m_2$ so that the total overlap integral for the case where both of the azimuthal quantum numbers are greater than or equal to zero looks like

$$\begin{aligned} S(\mathbf{R}) &= 8 \sum_{l=|l_1-l_2|,2}^{l_1+l_2} (i^{l_1}(i^{l_2 *})(i^{l*}) \left(\int k^2 dk f_1(k) f_2(k) j_l(kR) \right) \\ &\times \int d\Omega_k (Y_{l_1 m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{lm_a}(\mathbf{k}) \\ &\times (pref(l, m_a) P_{lm_a}(cos(\theta)) cos(m_a \phi)) \\ &+ (-1)^{m_1} Y_{l_1 -m_1}(\mathbf{k}) Y_{l_2 m_2}(\mathbf{k}) Y_{lm_b}(\mathbf{k}) \\ &\times (pref(l, m_b) P_{lm_b}(cos(\theta)) cos(m_b \phi))). \end{aligned} \tag{4.78}$$

4.9.3 Both m1 and m2 less than zero

The final case we have to consider involves products of spherical harmonics like

$$(Y_{l_1 m_1}(\mathbf{r}) - Y^*_{l_1 m_1}(\mathbf{r}))(Y_{l_2 m_2}(\mathbf{r}) - Y^*_{l_2 m_2}(\mathbf{r})) \tag{4.79}$$

and again we can just write down the different angular terms that will be produced by multiplying out the brackets;

$$Y_{l_1m_1}(\mathbf{r})Y_{l_2m_2}(\mathbf{r}) \rightarrow Y_{l_1m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_a}(\mathbf{k})Y^*_{lm_a}(\mathbf{R}). \tag{4.80}$$

$$-Y^*_{l_1m_1}(\mathbf{r})Y_{l_2m_2}(\mathbf{r}) \rightarrow -Y^*_{l_1m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_b}(\mathbf{k})Y^*_{lm_b}(\mathbf{R}). \tag{4.81}$$

$$-Y_{l_1m_1}(\mathbf{r})Y^*_{l_2m_2}(\mathbf{r}) \rightarrow -Y_{l_1m_1}(\mathbf{k})Y^*_{l_2m_2}(\mathbf{k})Y_{lm_c}(\mathbf{k})Y^*_{lm_c}(\mathbf{R}). \tag{4.82}$$

$$Y^*_{l_1m_1}(\mathbf{r})Y^*_{l_2m_2}(\mathbf{r}) \rightarrow Y^*_{l_1m_1}(\mathbf{k})Y^*_{l_2m_2}(\mathbf{k})Y_{lm_d}(\mathbf{k})Y^*_{lm_d}(\mathbf{R}). \tag{4.83}$$

Once again we pair off the angular terms appropriately and use the identity 4.62 in order to write the result in terms of real spherical harmonics.

$$\begin{aligned} I_1 &= Y_{l_1m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_a}(\mathbf{k}) \\ &\times \left(Y^*_{lm_a}(\mathbf{R}) + Y_{lm_a}(\mathbf{R})\right) \\ &- Y_{l_1-m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_b}(\mathbf{k}) \\ &\times (-1)^{m_1}\left(Y^*_{lm_b}(\mathbf{R}) + Y_{lm_b}(\mathbf{R})\right). \end{aligned} \tag{4.84}$$

The complete expression for the overlap integral in this particular case then looks like

$$\begin{aligned} S(\mathbf{R}) &= 8 \sum_{l=|l_1-l_2|}^{l_1+l_2} (i^{l_1}(i^{l_2*})(i^{l*}) \left(\int k^2 dk f_1(k) f_2(k) j_l(kR) \right) \\ S(\mathbf{R}) &= 8 \sum_{l=|l_1-l_2|,2}^{l_1+l_2} \left(i^{l_1}(i^{l_2*})(i^{l*}) \left(\int k^2 dk f_1(k) f_2(k) j_l(kR) \right) \right. \\ &\quad J \\ &\times (pref(l,m_a)P_{lm_a}(cos(\theta))cos(m_a\phi)) \\ &- (-1)^{m_1} y_{l_1-m_1}(\mathbf{k})Y_{l_2m_2}(\mathbf{k})Y_{lm_b}(\mathbf{k}) \\ &\times (pref(l,m_b)P_{lm_b}(cos(\theta))cos(m_b\phi))). \end{aligned} \tag{4.85}$$

We can see that the overlap integral between PAOs having m_1 and m_2 greater than or equal to zero and that between PAOs having both m_1 and m_2 less than zero is identical apart from a sign change of the second angular component. This completes our derivation of overlap integrals between PAOs involving real combinations of spherical harmonics.

4.10 Spherical Harmonic Triple Product

The integral over solid angle Ω in the expressions which we have developed for the overlap integrals (for example see equation 4.85 above) can be evaluated analytically, using the results of group theory [48]. In terms of Wigner 3-j symbols (which are explained in the reference given) the integral of the product of three spherical harmonics takes a particularly elegant form,

$$\begin{aligned} I_\Omega &= \int d\Omega Y_{l_1 m_1}(\Omega) Y_{l_2 m_2}(\Omega) Y_{l_3 m_3}(\Omega) \\ &= 4\pi \sqrt{\frac{(2l_1+1)(2l_2+1)(2l_3+1)}{(4\pi)^3}} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \begin{pmatrix} l_1 & l_2 & l_3 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned} \tag{4.86}$$

The Wigner 3-j symbols can themselves be expressed in terms of Clebsch-Gordan or vector coupling coefficients [49] [48] which can also be evaluated

analytically,

$$s^{l_1 l_2}_{L\mu\nu} = \frac{\sqrt{(L+l_1-l_2)!(L-l_1+l_2)!(l_1+l_2-L)!(L+\mu+\nu)!(L-\mu-\nu)!}}{\sqrt{(L+l_1+l_2+1)!(l_1-\mu)!(l_1+\mu)!(l_2-\nu)!(l_2+\nu)!}} \times \sum_x \frac{(-1)^{x+l_2+\nu}\sqrt{(2L+1)}(L+l_2+\nu-x)!(l-\mu+x)!}{(L-l_1+l_2-x)!(L+\mu+\nu-x)!x!(x+l_1-l_2-\mu-\nu)!} \tag{4.87}$$

The notation here is such that L = l_3, l_1 and l_2 are the three principal angular momentum indices, and $\mu = m_1$, $\nu = m_2$. The summation index, x, is restricted so that terms in the denominator are non negative. In terms of the vector coupling coefficient the Wigner 3-j symbol is

$$\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} = (-1)^{l_1-l_2-l_3}(-1)^{l_3-m_3}\frac{s^{l_1 l_2}_{l_3 m_1 m_2}}{\sqrt{2l_3+1}} \tag{4.88}$$

Thus by applying the formulae above one can evaluate the integral of a triple product of spherical harmonics.

4.11 Gradients of PAO Functions

This section gives the details of gradients of PAO functions (and thus gradients of overlap integrals since they have a similar functional form). As mentioned in section 4.2 these are required to evaluate the forces on the ions due to the electronic charge. There are only two distinct types of PAO function we must be concerned with - those having m less than zero and those having m greater than or equal to zero.

4.11.1 Spherical Coordinate System

Spherical coordinates are a popular choice of system for dealing with angular momentum, covariant spherical coordinates are defined by [49]

$$\begin{aligned} x_{+1} &= -\frac{1}{\sqrt{2}}(x+iy) = -\frac{1}{\sqrt{2}}r\sin(\theta)e^{i\phi} \\ x_0 &= z = r\cos(\theta) \\ x_{-1} &= \frac{1}{\sqrt{2}}(x-iy) = \frac{1}{\sqrt{2}}r\sin(\theta)e^{-i\phi}. \end{aligned} \tag{4.89}$$

Contravariant spherical coordinates are defined by

$$\begin{aligned} x^{+1} &= -\frac{1}{\sqrt{2}}(x-iy) = -\frac{1}{\sqrt{2}}r\sin(\theta)e^{-i\phi} \\ x^0 &= z = r\cos(\theta) \\ x^{-1} &= \frac{1}{\sqrt{2}}(x+iy) = \frac{1}{\sqrt{2}}r\sin(\theta)e^{i\phi}. \end{aligned} \tag{4.90}$$

We can write the differential operator ∇ in terms of the spherical basis vectors,

$$\begin{aligned} \nabla &= \sum_{\mu}(-1)^{\mu}\mathbf{e}_{\mu}\nabla_{-\mu} \\ \nabla &= -\mathbf{e}_{+1}\nabla_{-1} + \mathbf{e}_0\nabla_0 - \mathbf{e}_{-1}\nabla_{+1}. \end{aligned} \tag{4.91}$$

We can then write the spherical components of the gradient operator as

$$\begin{aligned} \nabla_{+1} &= -\frac{1}{\sqrt{2}}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right), \\ \nabla_0 &= \frac{\partial}{\partial z}, \\ \nabla_{-1} &= \frac{1}{\sqrt{2}}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right). \end{aligned} \tag{4.92}$$

Using spherical coordinates the different gradient components of a PAO (i.e. a radial function multiplied by a spherical harmonic) can be written as

$$\begin{aligned}\nabla_0\left(f(r)Y_{lm}(\theta,\phi)\right) &= \sqrt{\frac{(l+1)^2-m^2}{(2l+1)(2l+3)}}\left(\frac{df}{dr}-\frac{l}{r}f\right)Y_{l+1m}(\theta,\phi)\\ &+ \sqrt{\frac{l^2-m^2}{(2l-1)(2l+1)}}\left(\frac{df}{dr}+\frac{l+1}{r}f\right)Y_{l-1m}(\theta,\phi)\end{aligned} \tag{4.93}$$

and

$$\begin{aligned}\nabla_{\pm1}\left(f(r)Y_{lm}(\theta,\phi)\right) &= \sqrt{\frac{(l\pm m+1)(l\pm m+2)}{2(2l+1)(2l-1)}}\left(\frac{df}{dr}-\frac{l}{r}f\right)Y_{l+1m\pm1}(\theta,\phi)\\ &-\sqrt{\frac{(l\mp m-1)(l\mp m)}{2(2l-1)(2l+1)}}\left(\frac{df}{dr}+\frac{l+1}{r}f\right)Y_{l-1m\pm1}(\theta,\phi).\end{aligned} \tag{4.94}$$

So that the components of the gradient operator in spherical coordinates have a nice form when acting on a PAO function. If we look at the equations 4.92 then we see it is straightforward to form the Cartesian components of the gradient operator acting on a PAO function using by adding/subtracting the appropriate spherical components.

$$\partial x \sim \nabla_{+1} - \nabla_{-1} \tag{4.95}$$

$$\partial y \sim \nabla_{+1} + \nabla_{-1} \tag{4.96}$$

$$\partial z \sim \nabla_0. \tag{4.97}$$

This gives us a very convenient way of testing code written to calculate gradients of PAO functions, since we can express the PAOs analytically in

terms of Cartesian coordinates when the radial functions are Gaussian, and then we can also write down a full expression for the PAO gradients directly in Cartesian coordinates.

4.11.2 Application to CONQUEST PAOs

In CONQUEST we are using PAOs which involve real combinations of spherical harmonics and so when we take their gradient we end up with real combinations of spherical harmonics in the final expression. For example if we look at $\partial x(Y_{lm}(\hat{\phi}) + Y^*_{lm}(\hat{\phi}))$ where

$$Y_{lm}(\hat{\phi}) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}} e^{im\psi} P_{lm}(\cos\theta) \tag{4.98}$$

where $P_{lm}(x)$ is an associated Legendre polynomial. Forming the gradient from 4.95 above then gives us

$$\begin{aligned}
& -\sqrt{2}i\partial x f(r)(Y_{lm}(\theta,\phi) + Y^*_{lm}(\theta,\phi)) = P_1(l,m)F_1(f_r,l,m) \\
* \quad & (pref(l+1,m+1)P_{l+1m+1}(\cos(\theta))2\cos((m+1)\phi)) \\
- \quad & P_2(l,m)F_2(f_r,l,m) \\
* \quad & (pref_{l-1,m+1}P_{l-1m+1}(\cos(\theta))2\cos((m+1)\phi)) \\
- \quad & P_3(l,m)F_1(f_r,l,m) \\
* \quad & (pref_{l+1m-1}P_{l+1m-1}(\cos(\theta))2\cos((m-1)\phi)) \\
+ \quad & P_4(l,m)F_2(f_r,l,m) \\
* \quad & (pref_{l-1m-1}P_{l-1}m-1(\cos(\theta))2\cos((m-1)*\phi))
\end{aligned} \tag{4.99}$$

where pref is the prefactor for a complex spherical harmonic as in 4.98. The x component of the gradient for PAOs with m equal to zero is straightforward to work out, and the case where m is less than zero leads to post-multiplication by sines instead of cosines in the expression above. The pre-multipliers $P_1(l,m)$ through to $P_4(l,m)$ are

$$P_1(l,m) = \sqrt{\frac{(l+m+1)(l+m+2)}{2(2l+1)(2l+3)}} \tag{4.100}$$

$$P_2(l,m) = \sqrt{\frac{(l-m-1)(l-m)}{2(2l-1)(2l+1)}} \tag{4.101}$$

$$P_3(l,m) = \sqrt{\frac{(l-m+1)(l-m+2)}{2(2l+1)(2l+3)}} \tag{4.102}$$

$$P_4(l,m) = \sqrt{\frac{(l+m-1)(l+m)}{2(2l-1)(2l+1)}} \tag{4.103}$$

and the we have written the parts depending on the radial function as

$$F_1(f_r,l,r) = \left(f' - \frac{l}{r}f\right) \tag{4.104}$$

$$F_2(f_r,l,r) = \left(f' + \frac{(l+1)}{r}f\right) \tag{4.105}$$

Similarly the y component of the gradient of a PAO with m greater than zero

looks like

$$
\begin{aligned}
& -\sqrt{2}\partial y f(r)(Y_{lm}(\theta,\phi) + Y^*_{lm}(\theta,\phi)) = P_1(l,m)F_1(f_r,l,r) \\
* \quad & (pref(l+1,m+1)P_{l+1m+1}(\cos(\theta))2\sin(m+1(\phi))) \\
- \quad & P_2(l,m)F_2(f_r,l,r) \\
* \quad & (pref_{l-1,m+1}P_{l-1m+1}(\cos(\theta))2\sin((m+1)\phi)) \\
+ \quad & P_3(l,m)F_1(f_r,l,r) \\
* \quad & (pref_{l+1m-1}P_{l+1m-1}(\cos(\theta))2\sin((m-1)\phi)) \\
- \quad & P_4(l,m)F_2(f_r,l,r) \\
* \quad & (pref_{l-1m-1}P_{l-1}m-1(\cos(\theta))2\sin((m-1)*\phi)) \qquad (4.106)
\end{aligned}
$$

The y component of the gradient for PAOs with m less than zero again involves post-multiplying by cosine rather than sine functions. The z component of the gradient is straightforward to evaulate since it is effectively identical to the zeroth component of the gradient in spherical coordinates.

4.11.3 What happens when θ is nearly zero?

Originally when programming the gradients of PAO functions we wrote down the gradient operator in spherical polar coordinates and applied it directly to the PAO function, finally obtaining an expression,

$$
\begin{aligned}
\nabla f(r)C_{lm}P^l_m(cos(\theta))F(m\phi) \quad = \quad & (\sin(\theta)\cos(\phi)c_r + \cos(\theta)\cos(\phi)c_\theta - \sin(\phi)c_\phi)\,\mathbf{i} \\
+ \quad & (sin(\theta)\sin(\phi)c_r + \cos(\theta)\sin(\phi)c_\theta + \cos(\phi)c_\phi)\,\mathbf{j} \\
+ \quad & (\cos(\theta)c_r - \sin(\theta)c_\theta)\,\mathbf{k} \qquad (4.107)
\end{aligned}
$$

In the expression above

$$c_r = f'(r)C_{lm}P_m^l(\cos(\theta))F(m\phi). \tag{4.108}$$

$$c_\theta = \frac{1}{r}f(r)C_{lm}(-\sin(\theta))\left(\frac{1}{\sqrt{1-x^2}}P_{m+1}^l(x) - \frac{mx}{1-x^2}P_m^l(x)\right)F(m\phi). \tag{4.109}$$

$$c_\phi = \frac{1}{r\sin(\theta)}f(r)C_{lm}P_m^l(\cos(\theta))F'(m\phi). \tag{4.110}$$

where we have set $x = \cos(\theta)$ and $F(m\phi)$ is equal to $\cos(m\phi)$ if $m >= 0$ and $\sin(m\phi)$ otherwise. However when the polar angle θ is equal to zero and we are calculating gradients of our PAO function at some point on the z-axis the above expression becomes awkward since ϕ becomes ill-defined. The alternative form of the PAO gradient presented in the previous section is more useful in this case due to the identities

$$Y_{lm}(0,\phi) = \delta_{m0}\sqrt{\frac{2l+1}{4\pi}}. \tag{4.111}$$

$$Y_{lm}(\pi,\phi) = \delta_{m0}(-1)^l\sqrt{\frac{2l+1}{4\pi}}. \tag{4.112}$$

Because our formulation for the gradient of a PAO involves real combinations of spherical harmonics we can substitute the expressions above for our spherical harmonics whenever we wish to take the gradient at a point on the z axis.

4.11.4 PAOs having m less than zero

The results for applying the gradient operator to PAOs having m less than zero are very similar to those having m greater than zero, the difference

being a change of sign in the combination of spherical harmonics determining whether we get $\sin(m\phi)$ or $\cos(m\phi)$. The expressions look like

$$\begin{aligned}
&\partial_{x(y)}\left(f(r)(Y_{lm}(\mathbf{r}) - Y^*_{lm}(\mathbf{r}))\right) = P_1(l,m)F_1(f_r l,r,)(Y_{l+1m-1}(\mathbf{r}) \mp Y^*_{l+1m-1}(\mathbf{r})) \\
-\ & P_2(l,m)F_2(f_r,l,r)(Y_{l-1m+1}(\mathbf{r}) \mp Y^*_{l-1m+1}(\mathbf{r})) \\
+\ & P_3(l,m)F_1(f_r,l,r)(Y_{l+1m+1}(\mathbf{r}) \mp Y^*_{l+1m+1}(\mathbf{r})) \\
-\ & P_4(l,m)F_2(f_r,l,r)(Y_{l-1m-1}(\mathbf{r}) \mp Y^*_{l-1m-1}(\mathbf{r})). \qquad (4.113)
\end{aligned}$$

In the equation above the $\pm$ cases apply depending on whether we are taking the x(-) or y(+) component of the gradient. The premultipliers to the spherical harmonic combinations are as in 4.99 and 4.106 shown previously. Again the z component of the gradient is straightforward to calculate as it is equal to the ∇_0 term shown earlier.

4.12 Testing PAO Functions

Having coded up the overlap and kinetic energy integrals (related to the overlap integrals by a simple factor of k^2) between PAOs we found ourselves in a position to carry out numerical tests to verify the accuracy of the integrals done in k-space. One simple way to test the integration routines was to compare them with the analytic results computed when the PAO radial functions were Gaussians multiplied by r^l where the value of l is the orbital angular momentum of the PAO (this r^l prefactor is necessary to make sure the PAOs go smoothly to zero at the origin).

Again looking at the overlap integral of two arbitrary functions

$$S(R) = \int dr f_1^*(r) f_2(r'), \tag{4.114}$$

with $r' = r - R$, R being the displacement vector between the functions. We work with real PAOs in CONQUEST, so that the complex-conjugation of $f_1^*(r)$ can be ignored. Working with real PAO functions then allows us to express the PAOs in Cartesian coordinates as the product of an *lth* order polynomial in x, y, z with a Gaussian function appropriately normalised as described above.

In order to transform the PAOs into Cartesian coordinates we use the well-known identities;

$$\begin{aligned} x &= r\sin(\theta)\cos(\phi), \\ y &= r\sin(\theta)\sin(\phi), \\ z &= r\cos(\theta). \end{aligned} \tag{4.115}$$

It can then be shown that the real spherical harmonic combinations are easily expressible in terms of Cartesian coordinates.

4.12.1 Gaussian PAO Overlap Integrals

Here we show how to calculate the analytic expression for the (p,p) overlap integral, that is an overlap integral between PAOs each having orbital angular momentum l equal to one. There are three distinct cases which we can consider, corresponding to whether the PAO m values are both positive (a value of zero is taken here to be positive, since the rule for constructing a real

spherical harmonic for $m = 0$ is the same as for those having $m > 0$), both negative or one negative and one positive. This is because of the difference in the construction of real PAO combinations with m greater than or equal to zero as opposed to combinations with m less than zero. We will just evaluate the first type of integral, where both m values are greater than zero, as an illustration.

The real PAO is made by adding the complex PAO to its complex conjugate with normalisation factor $\frac{1}{\sqrt{2}}$. Our real PAO is given by $Y_1^1(\Omega)$ in 4.115 above. The radial function is chosen to be $f(r) = re^{-\alpha r^2}$, so that our PAO $F_1^1(r)$ is

$$F_1^1(r) = -\sqrt{\frac{3}{8\pi}} x e^{-\alpha r^2} \tag{4.116}$$

We are going to calculate the integral

$$\begin{aligned} S(R) &= \int F_1^1(r) F_1^1(r-R) dr \\ &= \int_{\infty}^{+\infty} (\frac{3}{8\pi}) x(x-X) e^{-\alpha r^2} e^{-\alpha(r-R)^2} \end{aligned} \tag{4.117}$$

expanding this out explicitly in the three Cartesian coordinates then gives,

$$S(R) = \frac{3}{8\pi} \int_{\infty}^{+\infty} dxdydzx(x-X) e^{-\alpha(x^2+y^2+z^2)} e^{-\alpha((x-X)^2+(y-Y)^2+(z-Z)^2)} \tag{4.118}$$

and we can simplify the exponentials to give

$$\begin{aligned} S(R) &= \frac{3}{8\pi} \int_{-\infty}^{\infty} dxdydz(x^2 - xX) e^{-2\alpha(x^2-xX)} e^{-2\alpha(y^2-yY)} \\ &* \quad e^{-2\alpha(z^2-zZ)} e^{-\alpha(X^2+Y^2+Z^2)}. \end{aligned} \tag{4.119}$$

So that the overlap integral has separated out into three independent integrals over the three Cartesian variables. We can evaluate

$$\begin{aligned} I_{y_0} &= \int_{-\infty}^{\infty} dy e^{-2\alpha(y^2 - yY)} \\ &= e^{\frac{\alpha Y^2}{2}} \sqrt{\frac{\pi}{2\alpha}}. \end{aligned} \tag{4.120}$$

where we find the result by completing the square and using standard identities. So $S(R)$ becomes

$$S(R) = I_{y_0} I_{z_0} \frac{3}{8\pi} \int_{-\infty}^{\infty} dx e^{\alpha R^2} (x^2 - xX) e^{-2\alpha(x^2 - xX)} \tag{4.121}$$

Again we use standard identities to evaluate the integral along the x axis, and

$$\begin{aligned} I_{x^2} &= \int_{-\infty}^{\infty} dx x^2 e^{-2\alpha(x^2 - xX)} \\ &= e^{\alpha \frac{X^2}{2}} \left(\frac{1}{4} \sqrt{\frac{\pi}{2\alpha^3}} + \frac{X^2}{4} \sqrt{\frac{\pi}{2\alpha}} \right). \end{aligned} \tag{4.122}$$

$$\begin{aligned} I_x &= \int_{-\infty}^{\infty} dx x e^{-2\alpha(x^2 - xX)} \\ &= \frac{X}{2} e^{\frac{\alpha(X^2)}{2}} \sqrt{\frac{\pi}{2\alpha}}. \end{aligned} \tag{4.123}$$

so that we have at last the overlap integral between two Y_1^1 Gaussian PAO functions,

$$S(R) = \frac{3\pi}{16\alpha} \left(\frac{1}{4} \sqrt{\frac{\pi}{2\alpha^3}} - \frac{X^2}{4} \sqrt{\frac{\pi}{2\alpha}} \right) e^{-\frac{\alpha}{2} R^2}. \tag{4.124}$$

This illustrates just one integral between the many different possible PAO angular momentum combinations. The integrals are exact and provide a very

useful and straightforward way of testing the effectiveness of the FFT based overlap integrals.

The results of tests for the different PAO combinations are tabulated below, and it is clear that the FFT based code agrees with the analytic calculation to within rounding error. All the comparisons were done with a value $\Delta k = 0.5$ and a k-space cut off of 10000. We note that the kinetic energy matrix elements are related to the overlap matrix elements by a factor of k^2 in k-space, and can be tested in the same way.

X	Y	Z	Analytic value	FFT value
0.1	0.1	0.1	5.3752185094753(475)E-002	5.3752185094753(315)E-002
0.2	0.2	0.2	4.91257982345664(96)E-002	4.91257982345664(33)E-002
0.3	0.3	0.3	4.22829663887394(07)E-002	4.22829663887394(42)E-002
1.0	1.0	1.0	2.7576650329348951E-003	2.7576650329348951E-003
2.0	2.0	2.0	3.4032290145(118220)E-007	3.4032290145(058885)E-007

Table 4.1: S-S Gaussian matrix elements

X	Y	Z	Analytic value	FFT value
0.1	0.1	0.1	1.02863614(09514178)E-004	1.02863614(10041409)E-004
0.2	0.2	0.2	3.76041059005(02054)E-004	3.76041059005(16007)E-004
0.3	0.3	0.3	7.282384626352(3037)E-004	7.282384626352(5035)E-004
1.0	1.0	1.0	5.277243915785(8289)E-004	5.277243915785(9132)E-004
2.0	2.0	2.0	2.60505455105(86556)E-007	2.60505455105(90024)E-007

Table 4.2: D(z2-r2)-D(yz) Gaussian matrix elements

4.13 PAO Force Test

Once the PAOs had been coded up and integrated into CONQUEST force tests were done using bulk Silicon, with one atom moved by 0.00054 Å along the x axis. The force on the atom in the y direction was calculated using both

the analytic formulae and by applying a finite difference approximation using the energy. Such a force is expected to be very small, and thus a stringent test of the numerical accuracy of the evaluation of the PAO matrix elements. Consistency of the two quantities demonstrates that the PAOs have been correctly implemented within the code as a whole, as well as the fact that they will behave correctly as a standalone code.

Force component	Numerical	Analytic
Total	-0.000000499512	-0.000000500998
Non-local	-0.000000339906	-0.000000339882
Local Hellmann-Feynman	-0.000000704448	-0.000000704749
Non-self-consistent	0.000001463663	0.000001462742
Kinetic phi Pulay	0.000000609113	0.000000609182
Local phi Pulay	-0.000001701750	-0.000001701737
S Pulay	0.000000363898	0.000000363933
Total Pulay	0.000005556799	0.000005556731

Table 4.3: Comparison of numerical and analytic PAO force components (Ha/Bohr).

The results of the test show that the PAOs have been correctly integrated into the CONQUEST code, and that the forces are of sufficient consistency with the energy to allow reliable structural relaxation calculations.

4.14 Conclusions

The success of SIESTA [10] using a basis of PAOs has led to their inclusion in CONQUEST too, and all the relevant technical details have been presented in this chapter. We have presented the full range of expressions, from the overlap integral between real PAO functions to their gradients, required for the implementation of a PAO basis set. We have also demonstrated the

accuracy of the code to evaluate matrix elements by using PAOs having Gaussian radial functions, whose overlap and kinetic energy integrals can be evaluated analytically.

In the preceding section tests comparing the forces found using finite differences in the total energy and the analytic force expressions in CONQUEST demonstrate that the code has been correctly integrated into CONQUEST. In the next chapter we will use PAOs generated by SIESTA to explore aspects of the linear scaling algorithm in CONQUEST on systems such as bulk Si and the Si (001) surface. The quick calculations afforded by the PAO basis should form a promising combination with the slower but systematically convergable blips.

Chapter 5

Silicon Tests

5.1 Introduction

In this chapter we perform tests of the different algorithms within CONQUEST on Silicon (bulk and (001) surface) with the newly implemented PAO basis. In order to generate the PAO functions themselves we have used the SIESTA Gen-Basis code [47]. The simulations discussed here have been performed using two processors in parallel.

The loop to find the total energy within CONQUEST can be run either self consistently (SC) using the Kohn-Sham (KS) energy functional or non self consistently (NSC) using the Harris-Foulkes (HF) energy functional. The solution of the KS equations may be done either using an order N (O(N)) method or by direct diagonalisation of the Hamiltonian matrix to find its eigenvectors, details of the order N method were presented in chapter three.

Here we begin by finding the ground state energy of bulk Si (64 atom unit

cell) using direct diagonalisation and comparing against the result obtained with the NSC Harris-Foulkes functional. We calculate strain curves of bulk Si using both single-zeta (SZ) and a double zeta plus polarisation (DZP) basis set, using PAOs with cut-off radii in the region of 5, 6, 7 and 8 Bohr.

Once we have performed comparisons of the NSC and SC functionals we then compare the accuracy of the O(N) algorithm against direct diagonalisation. The O(N) algorithm itself can also be used in either SC or NSC modes. The accuracy of the linear scaling result can be improved by increasing the range of the L matrix described in the previous chapter. As the L range is increased the localisation region becomes larger and larger until the calculated ground state energy converges towards the diagonalisation value with converged k point sampling. In order to demonstrate this convergence we calculate a series of strain curves using a SZ basis as a function of increasing L range and compare them to the strain curve obtained using diagonalisation and a 222 k point mesh (this k point sampling is enough to converge the diagonalisation result for our 64 atom test system see figure 5.5).

5.2 PAO Basis Functions

In figure 5.1 we have the radial functions of the PAOs comprising the single zeta basis set. The black curve shows the l = 0 (S) radial function, the red curve being the l = 1 (P) radial function. The radial functions are divided by r^l so the p radial functions do not go to zero at the origin but instead approach some constant value. In order to form the PAO functions we must multiply the radial part by a spherical harmonic. For the S function this is simply a constant, the real spherical harmonics associated with the P function

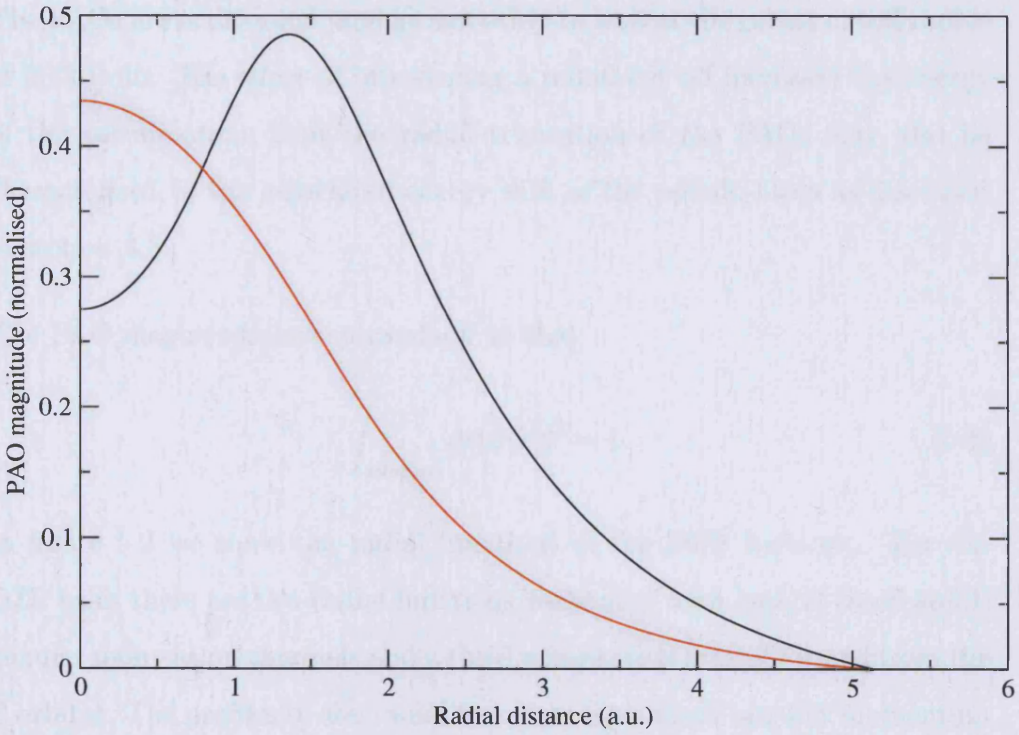

Figure 5.1: SZ PAO radial functions; S - black line, P - red line (R_{cut} = 5.13 Bohr).

are

$$\begin{aligned} Y_1^1(\Omega) &= \sqrt{\frac{3}{4\pi}}\frac{x}{r} \\ Y_1^0(\Omega) &= \sqrt{\frac{3}{2\pi}}\frac{z}{r} \\ Y_{-1}^1(\Omega) &= \sqrt{\frac{3}{4\pi}}\frac{y}{r} \end{aligned} \tag{5.1}$$

The PAOs are orthogonal, and go smoothly to zero at the preset cutoff radius of 5.13 Bohr. The effect of introducing a radial cut off increases the energy of the pseudo-atom, thus the radial truncation of the PAOs may also be characterised by the associated energy shift of the pseudo-atom as discussed in section 4.3.

The PAO magnitudes are normalised so that

$$\int_{allspace} dr|\psi(r)|^2 = 1. \tag{5.2}$$

In figure 5.2 we show the radial functions of the DZP basis set. For the DZP basis there are two radial functions associated with each of the S and P angular momentum channels and a third polarisation orbital formed from the P orbital. The profiles of the radial functions for a given angular momentum are quite similar because the value of the split radius is close to the radial cut-off. In order to generate these functions the SIESTA GenBasis code was used which also generated associated pseudopotentials using the Troullier-Martins method [50]. In the Troullier-Martins scheme the radial part of the pseudo wavefunction is defined according to Kerker's prescription [51], that is to be the all electron wavefunction outside a chosen core radius, and a

parameterized expression within it,

$$R_l^{PP}(r) = R_l^{ae}(r) \quad (r \geq r_c) \tag{5.3}$$

$$R_l^{PP}(r) = r^l exp[p(r)] \quad (r \leq r_c) \tag{5.4}$$

Troullier and Martins proposed their own form for the polynomial $p(r)$ and give the criteria for determining the necessary coefficients in [50]. Their scheme is a popular one for generating smooth and transferable pseudopotentials.

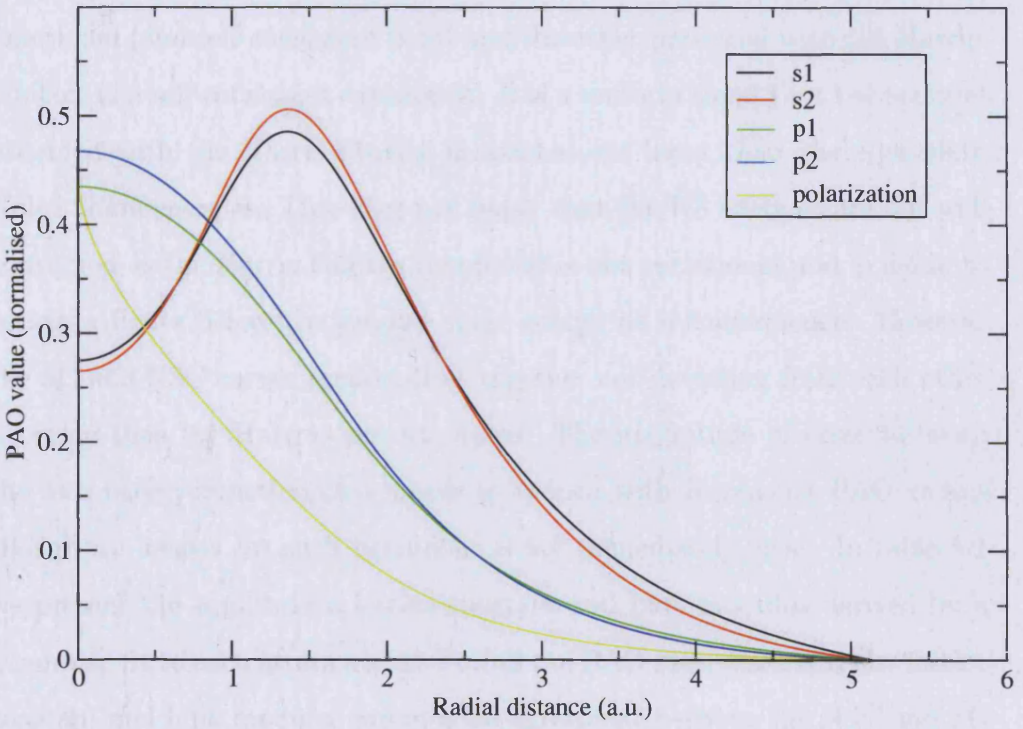

Figure 5.2: Radial functions of DZP PAOs (5.13 Bohr); S - (black, red), P - (green, blue), Polarisation - (yellow).

5.3 Results of Diagonalisation Tests

We produced a series of strain curves using PAOs having cut-off radii of 5.13, 5.96, 6.93 and 8.05 Bohr with both SZ and DZP basis sets in order to compare the accuracy of the NSC energy functional against the SC energy. In figure 5.3 we present results obtained with the single-zeta basis set. As the PAO radius is increased from 5.13 Bohr (through 5.96, 6.93) to 8.05 Bohr the total energy of the bulk is lowered though we note that the total energy is not necessarily variational with respect to PAO radius. For each PAO radius we have two strain curves, one produced using the Kohn-Sham energy functional (and self-consistent field) and the other produced with the Harris-Foulkes non self consistent expression. It is a uniform trend that the energies obtained with the Harris-Foulkes functional are lower than the equivalent Kohn-Sham energies. This does not imply that the KS energies are not well converged as the Harris-Foulkes functional is not variational and is liable to return a figure below the ground state energy as a consequence. However the SC and NSC curves remain close together not deviating from each other by more than 0.1 Hartree per 64 atoms. The magnitude of error between the two energy functionals appears to reduce with increasing PAO radius though the reason for such behaviour is not immediately clear. In table 5.1 we present the equilibrium lattice constant and bulk modulus derived by a quadratic fit to each strain curve. For all the PAO radii discussed the lattice constant and bulk modulus are in good agreement between the NSC and SC functionals, with errors in the bulk modulus at less than 5%.

In figure 5.4 we show the strain curves obtained using the DZP basis, again with PAO lengths varying from 5.13 up to 8.05 Bohr. The distribution of the curves is very different to that in figure 5.3 apart from the curves for

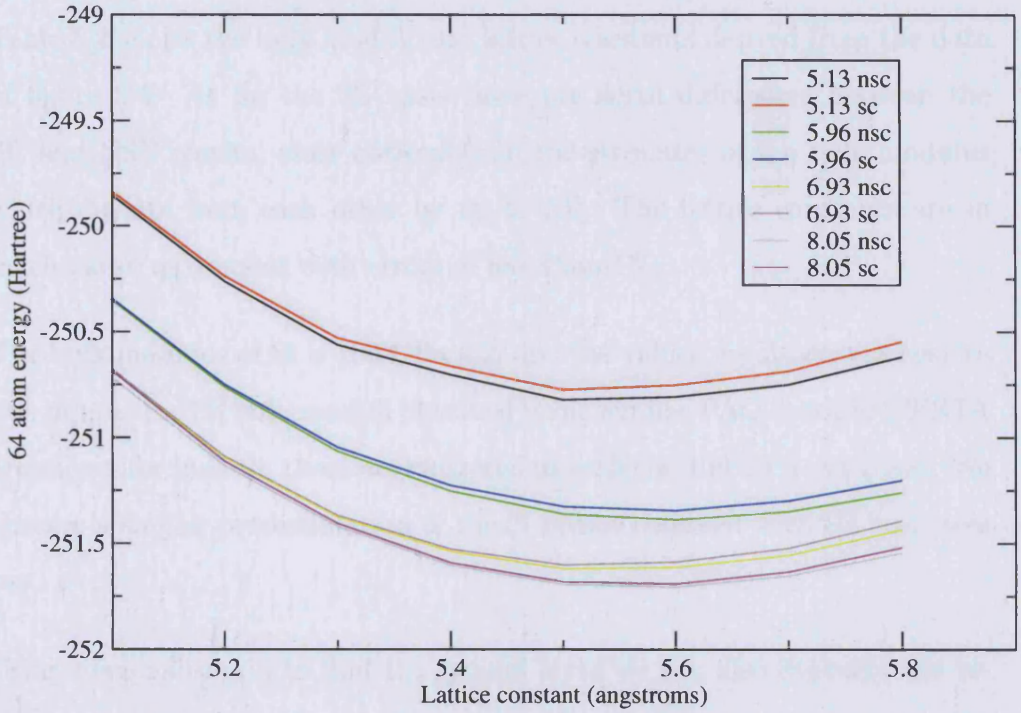

Figure 5.3: SZ Strain curves obtained via direct diagonalisation (gamma point); red 5.13 Bohr SC, black 5.13 Bohr NSC, blue 5.96 Bohr SC, green 5.96 Bohr NSC, brown 6.93 Bohr SC, yellow 6.93 Bohr NSC, purple 8.05 Bohr SC, grey 8.05 Bohr NSC.

the 5.13 Bohr PAO the rest are close together. However the (SC) energy obtained using the 6 Bohr PAOs is lower than that obtained with the 7 or 8 Bohr orbitals. This shows that the energy is not variational with respect to the length of the PAOs we have chosen. A trend also observed for the SZ basis is that the NSC energies are lower than the SC energies. Again this is a reflection of the non-variational nature of the Harris-Foulkes functional.

Table 5.2 shows the bulk moduli and lattice constants derived from the data of figure 5.4. As for the SZ basis there are small differences between the SC and NSC results, most noticeably in the estimates of the bulk modulus which deviate from each other by up to 5%. The lattice constants are in much closer agreement with errors of less than 1%.

The bulk modulus of Si is 100 GPa [52] and the values we obtain are near to this figure. In [10] bulk moduli obtained using similar PAO bases in SIESTA are shown for bulk Si, these are clustered around the 100 GPa mark too. We observe a similar overestimation of the Si lattice constant with SZ basis sets too.

Using diagonalisation to find the ground state we can also converge the results with respect to k point sampling. This will be important when we compare diagonalisation energies against order N energies. Figure 5.5 shows energies calculated with a 6 Bohr PAO (SZ) at gamma point, 222 and 444 k point sampling. Though there is a 0.25 Hartree lowering of the energy when the sampling is increased from gamma point to a 222 k point mesh the energies are well converged with the 222 mesh, as there is a barely observable difference with the 444 k point mesh. Thus energies obtained by direct diagonalisation with a 222 k point mesh will be sufficiently accurate to allow comparison against the order N algorithm.

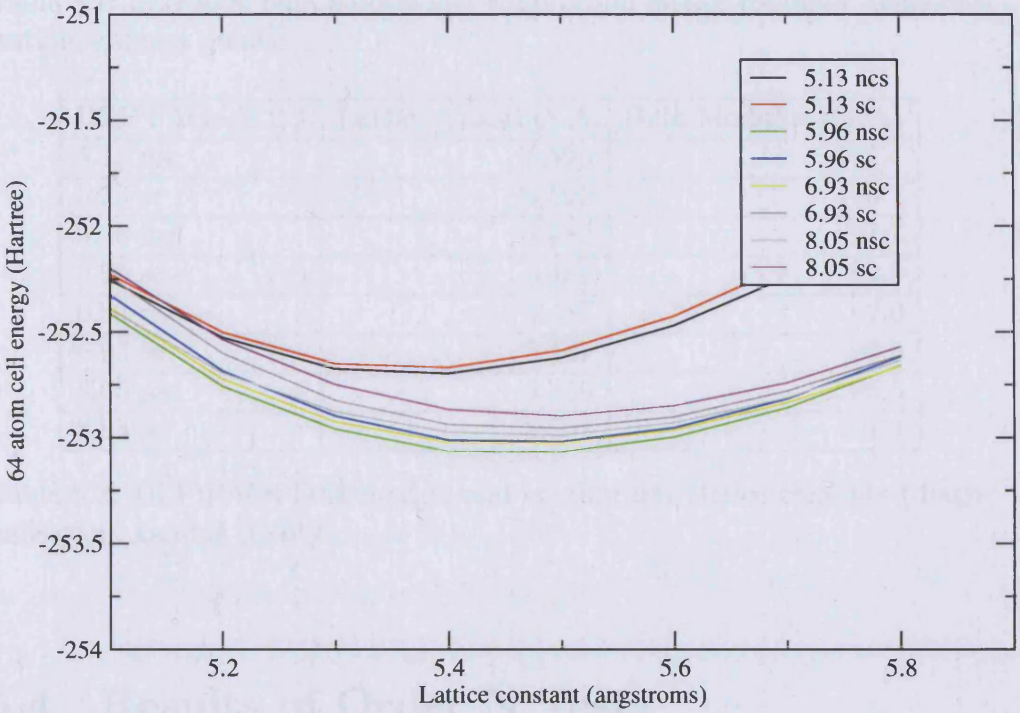

Figure 5.4: DZP strain curves by diagonalisation (gamma point). (R_c in a.u.)

SZ PAO radius (a.u.)	Lattice Constant Å	Bulk Modulus (GPa)
5.13 nsc	5.534	96.4
5.13 sc	5.537	96.6
5.96 nsc	5.597	90.8
5.96 sc	5.588	91.2
6.93 nsc	5.570	93.9
6.93 sc	5.557	96.1
8.05 nsc	5.579	93.5
8.05 sc	5.576	96.8

Table 5.1: SZ PAOs bulk moduli and equilibrium lattice contants (diagonalisation, gamma point).

DZP PAOs (a.u.)	Lattice Constant Å	Bulk Modulus (GPa)
5.13 nsc	5.396	103.4
5.13 sc	5.392	106.7
5.96 nsc	5.482	97.7
5.96 sc	5.488	99.9
6.93 nsc	5.487	92.0
6.93 sc	5.487	96.0
8.05 nsc	5.506	85.5
8.05 sc	5.501	91.4

Table 5.2: DZP PAOs bulk moduli and equilibrium lattice contants (diagonalisation, gamma point).

5.4 Results of Order N Tests

The linear scaling algorithm of CONQUEST relies on localisation of the single particle density matrix to make the density matrix sparse so that linear scaling sparse matrix multiplication techniques can be used. The auxiliary density matrix cut off can be fixed through the range L matrix described in the previous section. As the L matrix is made larger and larger the computation becomes more accurate and should eventually converge to the diago-

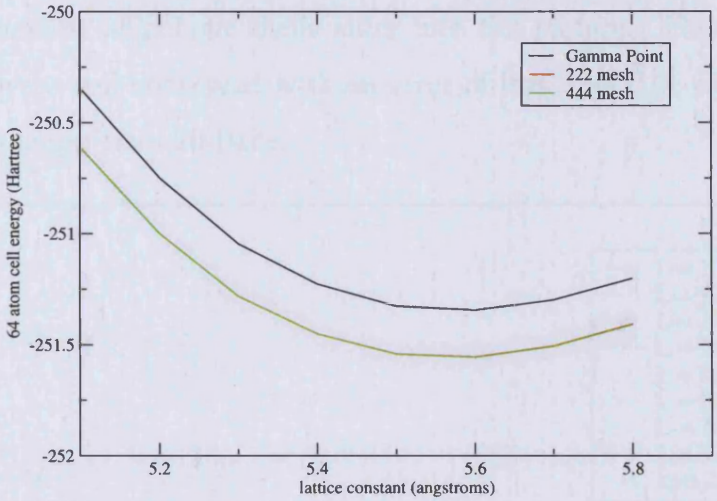

Figure 5.5: Comparison of K point results for 6 Bohr PAO (SZ, SC).

nalisation result. The cost of the computation does however scale as L_{cut}^2 as the number of L matrix elements increases.

We performed tests showing the convergence of linear scaling energies towards the diagonalisation values by producing series of strain curves at successively larger values of L_{cut}. The tests were done with a SZ basis, using both 5 and 6 Bohr PAOs. The results for the 5 Bohr PAO are plotted in figure 5.6. There is a large drop of about 0.75 eV in energies going from L_{cut} = 10 to L_{cut} = 12. This may be due to the increased interaction range picking up the effects of second nearest neighbour atoms which are 7.26 Bohr away. The third nearest neighbour shell is at 8.51 Bohr and we see that the strain curves are well converged only after L_{cut} = 20. The diagonalisation curve at 222 k points is shown on the graph and is very close to the curves at L_{cut} = 20 or more, demonstrating correct convergence of the order N algorithm.

Figure 5.7 shows a similar set of results developed using 6 Bohr PAOs in a SZ basis. Similar drops in energy to those seen in 5.6 are observed as more

and more nearest neighbour shells enter into the picture. The results are again relatively well converged with an error of less than 0.05 eV per atom once L_{cut} is larger than 20 Bohr.

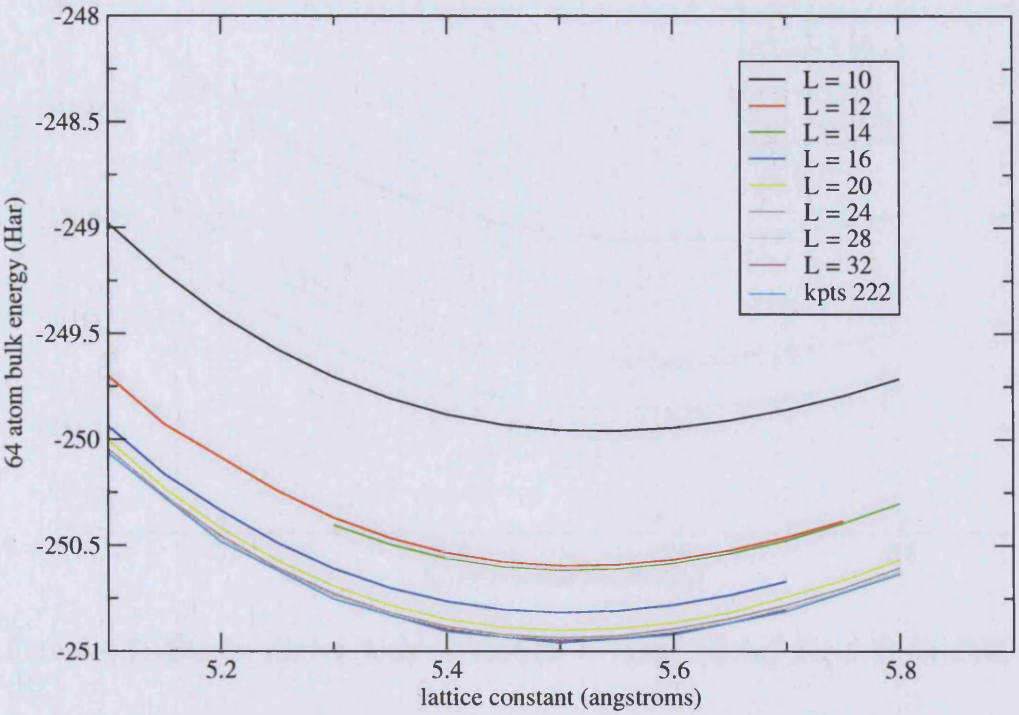

Figure 5.6: Strain curves with increasing L range (Bohr) for 5 Bohr PAO (SZ). .

In table 5.3 we have shown the bulk moduli and lattice constants derived from figure 5.7 using a quadratic fit. The lattice constant remains fairly stable as the L range is varied, but the bulk modulus varies by as much as 10% showing the latter quantity to have more sensitivity with respect to this parameter. Since the bulk modulus is found using the second derivative of the energy w.r.t. the cubic lattice constant it is more sensitive to small changes in the strain energy curve. Finally in figure 5.8 we plot the energy at a single value of the lattice constant (5.5 Å) as a function of L range (5 Bohr

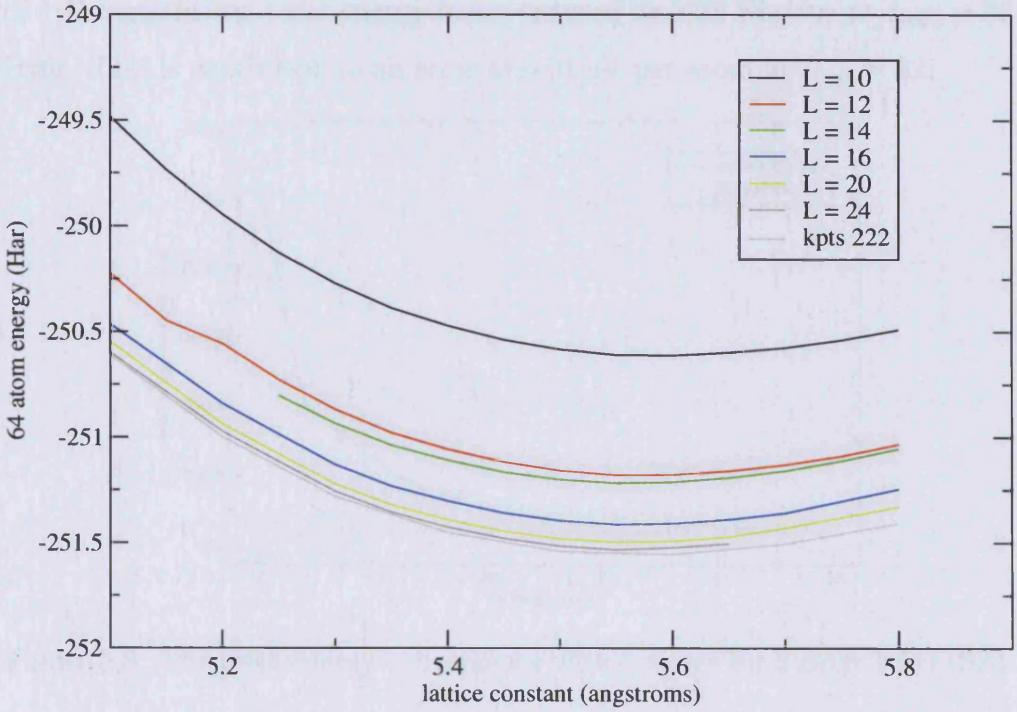

Figure 5.7: Strain curves with increasing L range (Bohr) for 6 Bohr PAO (SZ) .

L range (a.u.)	bulk modulus (GPa)	lattice constant Å
10	103.9	5.55
12	106.7	5.53
16	112.5	5.51
20	107.7	5.52
24	106.8	5.52
28	101.6	5.52
exact diagonalisation (222 kpts)	106.9	5.53

Table 5.3: Bulk moduli and equilibrium lattice constants (6 Bohr PAO, increasing L range).

PAO SZ basis set). The convergence appears to be exponential with an error of 1 Hartree in the total energy being reduced to 0.02 Hartree at $L_{cut} = 32$ Bohr. This is equivalent to an error of 0.01 eV per atom at $L_{cut} = 32$.

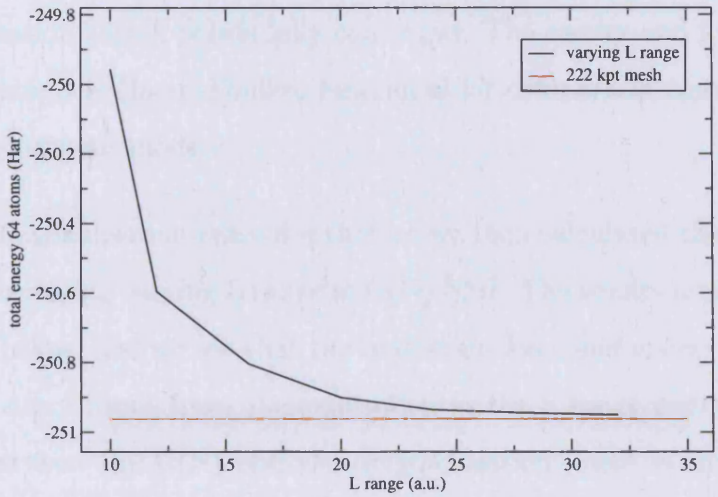

Figure 5.8: The total energy convergence with L range for 5 Bohr PAO (SZ).

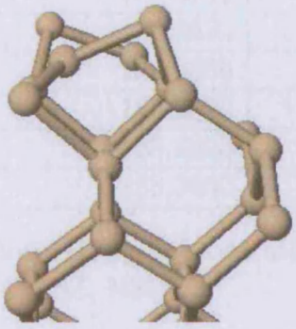

Figure 5.9: A segment of the Si (001) surface.

We also performed convergence tests of the linear scaling solution on the Si (001) surface. A segment of the surface is shown in figure 5.9, the characteristic dimers can be seen at the top. A surface cell containing 48 atoms was used for testing. The surface dimers were relaxed using diagonalisation

of the KS Hamiltonian and a SZ basis having a 6 Bohr radial cut off. The energies and geometry obtained from a 221 k point mesh were compared against those from a 441 k point mesh and found to be converged, this is important when testing the O(N) algorithm which gives results comparable to diagonalisation with k points fully converged. The energy and forces were calculated using the Harris-Foulkes functional for comparison against O(N) in non self consistent mode.

Taking the diagonalisation relaxed structure we then calculated the total energy and forces using varying L range in O(N) NSC. The results are displayed in table 5.4 below, and we see that the maximum force and energy converge towards the exact result from diagonalisation as the L range increases. The difference between the O(N) and the diagonalisation result is small but a relatively large L range of 32 Bohr is required to obtain such agreement.

L range (a.u.)	48 atom Total Energy (Har)	Max Force (Har/Bohr)
10	-187.6022	0.0047
12	-188.0178	0.0035
14	-188.0780	0.0027
20	-188.2653	0.0013
24	-188.2896	0.0011
28	-188.2982	0.0010
32	-188.3023	0.0009
2x2x1 kpts	-188.3060	0.0006

Table 5.4: Table showing the total energy of a 48 atom Si (001) surface slab and maximum force (6Bohr PAO SZ, NSC)

5.5 Conclusions

In this section we have performed tests of different functionalities within CONQUEST using the PAO basis set whose implementation was described in the previous chapter. We compare the energies obtained for a 64 atom cell of bulk silicon using diagonalisation of the Hamiltonian matrix within both self consistent and non self consistent modes and the results are found to be in good agreement showing the Harris-Foulkes functional to be a good approximation in bulk silicon. We also compare the performance of the linear scaling algorithm against direct diagonalisation. The order N results are found to be improvable by increasing the range of the L matrix discussed previously and figure 5.8 shows the exponential convergence of the total energy towards the exact diagonalisation value. However the convergence of the bulk modulus appears to be less stable with respect to the L range, perhaps as this quantity is more sensitive to the shape of the strain curve than for example the equilibrium lattice constant.

Using a Si(001) surface slab the total energy and maximum force within the O(N) NSC is shown to converge towards the value obtained by diagonalisation (NSC) using a 6 Bohr SZ basis set, with an L range of 32 Bohr the agreement between the O(N) and diagonalisation force and energy for the Si (001) is very good.

Chapter 6

Strained Growth of InAs on GaAs(110)

6.1 Introduction

In this chapter we discuss studies of misfit dislocation formation during the strained growth of Indium Arsenide (InAs) on the (110) surface of Gallium Arsenide (GaAs). Though both semiconductors have the zinc-blende lattice structure the lattice constant of InAs (6.06 Å) is seven percent larger than that of GaAs (5.65 Å). Thus the deposited InAs undergoes compressive strain leading to an increase in its internal energy. Eventually as the InAs coverage increases the strain energy becomes sufficient for the InAs to deform plastically with the onset of a strain relieving edge dislocation network. Our calculations are performed using VASP which is a conventional plane wave DFT code.

These observations of plastic relaxation were made during experiments performed by Belk et al ([16],[12]), in which InAs was grown on top of GaAs substrate using molecular beam epitaxy (MBE). The properties of the resulting crystal were monitored using a combination of scanning tunnelling microscopy (STM) imaging and in situ reflection high energy electron diffraction (RHEED) measurements, revealing the formation of a strain relieving edge dislocation network by three epilayers of coverage (number of layers denoted by θ, units monolayers, ML).

Belk et al. found that the first two layers of InAs deposited grew directly onto the underlying substrate without plastic deformation. By two epilayers coverage the compressive strain energy was not enough to cause any dislocation formation in the InAs lattice. However after deposition of the third epilayer STM and transmission electron microscopy (TEM) images revealed the presence of misfit dislocations along the [001] direction, these pure edge dislocations had Burgers vectors [1] of $(a_0/2)[1\bar{1}0]$, relieving the compressive strain along $[1\bar{1}0]$, though residual strain remained along [001] (see figure 6.1). Due to the lack of a threading component linking the misfit dislocations to the (110) surface Belk et al conclude that the dislocations must form in the second epilayer or above as these layers lie immediately below the growth surface.

Studies of this system have also been performed by Oyama et al[53][54], who repeat the experiment of Belk et al and also provide DFT calculations of the predicted core structure and critical epilayer thickness (θ_{crit}) at which onset of edge dislocations occurs. However they do not provide a direct

[1]The Burgers vector is used to characterise dislocation geometry. It is defined as the difference between a loop taken around the dislocation core in the bulk crystal and a loop in the bulk taken without the presence of the dislocation.

calculation of the critical thickness but instead interpolate between their DFT energies using a classical expression for the free energy. A study has also been performed by Maroudas et al [55] in which a mean field theory for the strain relaxation due to misfit dislocations is applied to reproduce the experimental results of [16].

Belk et al assert that the dislocations form in the second layer or above due to their appearance at $\theta = 3$ but Oyama et al believe the dislocations to be located at the heterointerface itself. The goal of our study is to calculate and compare the energies of dislocation networks at both of these positions in order to obtain the energetically favoured location.

The structure of this chapter is as follows; in section 6.2 we perform a literature review of work on the strained InAs/GaAs(110) heteroepitaxy. Section 6.3 presents details of our calculations on the bulk semiconductors, including technical convergences. We then compare the results (lattice constants and cohesive energies) obtained against the experimental values in sections 6.3.7, 6.3.8. Particularly important in understanding the misfit dislocations are the properties of InAs under the compressive strain it experiences on top the GaAs substrate. Details of the change in cohesive energy of InAs under both biaxial ([001] and [1$\bar{1}$0]) and uniaxial ([001]) strain are in sections 6.3.9, 6.3.10. These energies will be important in the later analysis of misfit dislocations.

Having performed calculations on the bulk semiconductors we turn our attention to the (110) surfaces, which have been extensively treated in the previous literature, in section 6.4. Firstly we review previous work in section 6.4.1 before discussing our own technical convergences and surface energies in sections 6.4.2, 6.4.6 and 6.4.7.

An examination of the growth of coherent epilayers of InAs on GaAs(110) forms section 6.5.3 and precedes the presentation of our calculations on the misfit dislocation characteristics in 6.5.6. We establish the position of the lowest energy dislocation core in section 6.5.9 and proceed to calculate the critical epilayer thickness for plastic deformation in section 6.5.11.

6.2 Literature Review

Here we discuss the work of Belk et al. and Oyama et al. in more detail as an understanding of the experimental conditions and data is important in producing ab initio results that provide more information at the atomistic level.

Belk et al [16] grew the semiconductor crystal using MBE in which solid sources are evaporated in an ultra-high vacuum (UHV) chamber. The molecular beams produced are projected onto a heated substrate surface at which crystallisation of the beam compound occurs. Within the MBE chamber electron diffraction can be used to monitor the growth progress of the crystal, allowing the rate of epitaxy to be measured and thus controlled. Belk et al use reflection high energy electron diffraction (RHEED) which is a popular technique for this purpose because the electron gun and screen set-up do not interfere with the solid source beams due to their orientation. During RHEED monitoring a high energy electron beam is directed at a slight grazing angle to the crystal surface (typically $1-3°$). The electrons are then scattered by the surface and form a diffraction pattern corresponding to the surface reciprocal lattice. The shallow incidence of the electron beam onto the crystal means that it does not interfere with MBE source beams and

also that the electrons do not penetrate far into the substrate, being principally scattered by the surface. Thus resulting diffraction patterns reflect the geometry of the growth surface rather than substrate lattice underneath.

Belk et al also produce STM images of different stages of InAs coverage providing "snapshots" of the system as a function of coverage. The STM images are formed by monitoring the tunnelling current between an atomically sharp tip and the sample surface placed a few atomic units away. A voltage bias is applied between tip and sample to induce the quantum mechanical tunnelling current. STM images can be produced in either of two modes, filled states or empty states, the distinction being the polarity of the applied voltage between the STM tip and sample. When the sample is negatively biased relative to the tip imaging is in filled states mode. The STM images produced by Belk et.al. are taken in filled states mode, and show Arsenic atoms at the growth surface rather than the group III atoms.

During the epitaxial growth strain relief was observed to occur through formation of two different dislocation networks. The first network formed consisted of ideal edge dislocations with line vector along [001], and a Burgers vector of $(a_0/2)[1\bar{1}0]$ as shown in figure 6.1, the second network being composed of 60 degree type dislocations rather than pure edge types. Complete strain relief in the biaxially compressed InAs is achieved only through both networks, as the edge dislocations effect strain relief along $[1\bar{1}0]$ only, having no effect on the compressive strain along [001]. We do not attempt a study of the 60 degree type dislocations as the network forms only after many tens of InAs epilayers have been deposited, and the computational expense of studying such a system is too large. By contrast the ideal edge dislocations form at much lower InAs coverage, with a fully formed network observed

after deposition of the fifth epilayer.

Belk et al observe that the initial deposition of the first layer of InAs is practically identical to the same stage of GaAs homoepitaxy, the compressive strain has no observable effect on the growth surface. As the first layer is deposited the InAs forms a number of separate 2D nuclei. The nuclei have heights of one or two layers in total, and are a common feature of (110) epitaxial growth. InAs has a lower heat of formation than GaAs so that the formation of surface alloys of (In,Ga)As is not favoured, Belk et. al. estimate the amount of In alloys by substitution into the GaAs substrate using observed STM contrast between the Ga and In atoms to be 1% of the surface (In) metal population, confirming that surface alloying is rare.

Eventually deposition of InAs leads to different surface characteristics compared to GaAs homoepitaxy. By $\theta = 0.75$ the island nuclei have lost any bilayer sections previously contained and become separated by monolayer trenches along [001] with a measured width of 6 Å. The spacing between trenches is estimated at 30 Å. The trenches become a regular feature of the growth surface at $\theta = 2$, and the islands form a "crazy paving" type structure atop the substrate, consisting of a network of irregular two dimensional islands. If the growth is carried out at a higher temperature of 480°C then this crazy paving is not seen, and instead a pattern of evenly spaced monolayer trenches are observed lying along the [001] direction, with a spacing of about 60 angstroms.

Plastic relaxation of the surface occurs at $\theta = 3$, the small islands begin to coalesce rather than remain separated by monolayer trenches, but contain elongated surface depressions of about 0.5 Å in the [001] direction, the depth of the depressions is much less than the monolayer step height, suggesting

that they are not trenches but are indicative of the presence of underlying edge dislocations. RHEED and TEM measurements of the system also show significant strain relaxation of the lattice along the $[1\bar{1}0]$ direction, confirming the presence of an array of misfit dislocations.

The lattice parameter of the epilayer is charted from $\theta = 1$ to 5 by RHEED, which shows it expanding towards that of InAs after three epilayers have been deposited. By $\theta = 5$ the surface depressions / edge dislocations form a periodic array in the [001] direction with an average spacing of sixty angstroms. Belk suggests that the formation of the edge dislocations from the crazy-paving islands places them at the second or third InAs epilayers, since the surface dips are first observed at $\theta = 3$ to 4, and not at the first epilayer supposing the dislocations to lie directly below the atomic planes at which surface dips are first observed.

Belk et al [12] also produce some data on the surface strain field due to the dislocation core, including the vertical surface displacement, the full width at half maximum (FWHM) and the dislocation spacing. They measure the depth of the surface dips as 0.7 Å (estimating the vertical resolution of the STM to be 0.1 Å) at both $\theta = 3$ and $\theta = 5$. The FWHM of the surface dips is measured as 15 Å at these two values of θ, before eventually reaching an upper limit of 35 Å by $\theta = 30$. We note that they claim a depth of 0.5 Å for the surface dips in the paper [16], but this is changed to 0.7 Å in the PhD thesis of Belk which was written after the paper and presents more detailed results on the system.

Belk's work in characterising the heteroepitaxial growth of InAs on GaAs(110) was reproduced by Oyama et al [53],[54], who studied the system experimentally and also computationally with density functional theory. Experimen-

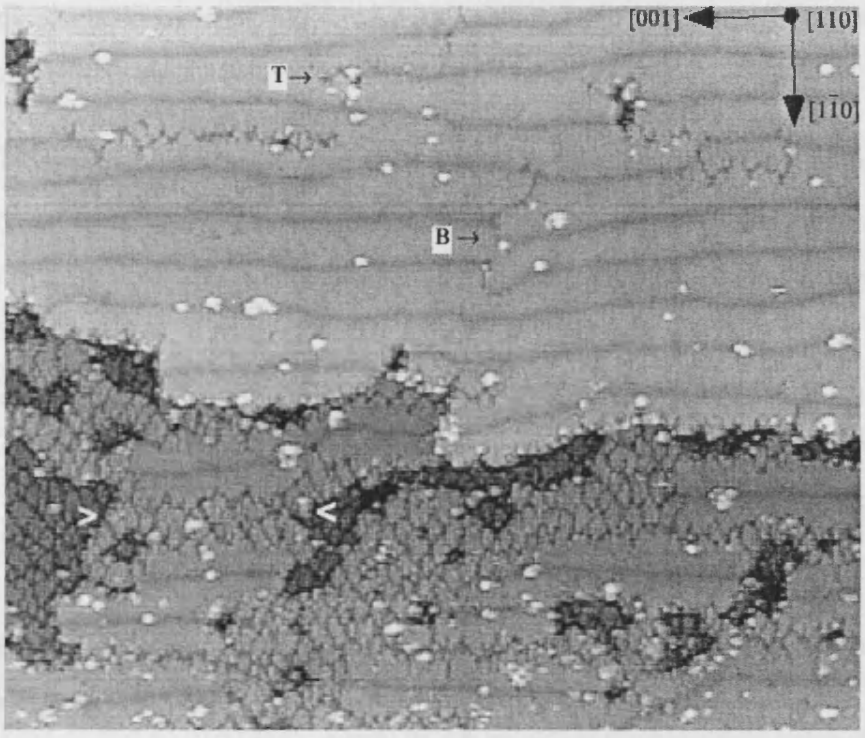

Figure 6.1: STM of misfit dislocations at 5 InAs epilayers (dark depressions along [001]).

tally they too observe the formation of an edge dislocation network at an early stage of InAs deposition proceeding to obtain a dislocation core structure through complementary atomistic calculations. In contrast with Belk et al, they believe the network to lie at the first layer of InAs at the heterointerface not the second layer above it. They present less detail on the growth features at different InAs coverages than Belk et al, but observe similar surface dips spaced by sixty angstroms along the [001] direction, corresponding to the misfit dislocations.

In a later paper Okajima et al [56] proceed to develop a phenomenological theory, placing the critical epilayer thickness θ_{crit} required for dislocation formation at 2.35 ML. Using an energy balance equation which is based on classical elasticity theory (Frank-Van der Merwe) the authors write down the expression for θ_{crit} in terms of the dislocation formation energy and the effective elastic constant of the strained growth system. They then fit the data from the DFT calculations of Oyama et al [53] to provide unknown parameters in the equations.

Okajima et al. [56] write down the free energy of the 2D coherent growth mode as

$$E_{coh}^{2d} = \gamma + \frac{1}{2}\hat{M}\epsilon_0^2 h. \tag{6.1}$$

where γ is the epilayer surface energy, ϵ_0 is the effective strain and $\hat{M}$ the effective elastic constant,

$$\hat{M} = \frac{2\mu(1+\nu)}{(1-\nu)}. \tag{6.2}$$

Here μ is the shear modulus and ν represents Poisson's ratio. The variable h represents the height of the strained layers. For the case of 2d incoherent

growth involving a MD network the expression is more elaborate;

$$E_{MD}^{2d}(l,h) = \gamma + \frac{1}{2}\hat{M}\left(\epsilon_0(1-\frac{l_0}{l})\right)^2 h + \frac{E_d}{l} \tag{6.3}$$

where l is the average dislocation spacing, l_0 is the ideal dislocation separation and E_d is the dislocation formation energy. By subtracting these two equations from one another and equating the difference to zero one can gain an estimate of the critical epilayer thickness θ_{crit},

$$\Delta E(h) = \frac{E_d}{l_0} - \frac{1}{2}\hat{M}\epsilon_0^2 h. \tag{6.4}$$

the energy difference $\Delta E(h)$ has been expressed as a function of height so that the critical thickness, i.e. the value of h at which $\Delta E(h)$ is zero, can be found.

The authors calculate the parameters for the cases of two InAs epilayers and four InAs epilayers (on a total of four GaAs substrate layers with the bottom layer held in fixed bulk positions and hydrogen terminated). They find $\Delta E(h)$ of 3.15 eV /cell and 0.85 eV/cell respectively ([53] [56]) - though they do not indicate of what corrections made to the cell energies for the different numbers of atoms present in the different structures. It is strange that the energy difference is larger for the smaller number of epilayers. It is also unusual that they did not perform the energy comparison at $\theta = 3$ which is where the dislocations are first seen to appear experimentally. We also note that the dislocation core structure appearing in [54] and [56] is asymmetric indicating that there are residual forces present in the system which will distort the energy comparisons which are made. We also note here an error in equation 6.4 which comes from equating the dislocation formation energy

with the coherent strain energy term. The pseudomorphic InAs(110) surface will have a different lattice constant in the $[1\bar{1}0]$ direction than the surface above the dislocation network and so the two surfaces will also have different associated energies. The fact that the two γ's in 6.1 and 6.3 are not equal has been neglected in the calculation of Oyama et al, though an estimate of the difference using DFT is possible. Using the formalism above Okajima et al gain $\theta_{crit} = 2.35$ and state that the first dislocations should be seen to form after deposition of three InAs epilayers, which agrees with their experimental results as well as those of Belk et al [16].

6.3 Bulk Calculations

Here we will examine basic physical properties of InAs and GaAs as a good understanding of these will be necessary later in our more complex studies of the strained growth system. We perform calculations on the bulk crystals and also on the (110) surfaces (in section 6.4). We calculate the cohesive energy and theoretical lattice constants of the bulk semiconductors and compare with experiment.

In our studies of InAs we also focus on the effects of the compressive strain it undergoes when deposited directly onto the GaAs substrate. In the initial stages of deposition the InAs is free to accomodate the misfit strain only through vertical expansion, being compressed in the (110) plane until at a critical thickness a dislocation network forms leaving residual strain along [001] only. Thus we model the behaviour of bulk InAs under both biaxial strain in (110) and when strained uniaxially along [001]. We calculate the change in the cohesive energy and also the equilibrium (110) interlayer

spacing under the different strain conditions, finding increase in the (110) interlayer spacing for both strains, as well as a decrease in the cohesive energy, which we would expect according to classical elasticity theory. The (110) spacing for biaxially strained InAs is greater than that under uniaxial strain.

6.3.1 Physical Properties of the Bulk Semiconductors

Both GaAs and InAs have the same zinc-blende crystal struture, consisting of two interpenetrating FCC lattices related by $(\frac{1}{4}\frac{1}{4}\frac{1}{4})$ along [111]. One fcc lattice corresponds to the group III species, the other to group V. The zinc-blende structure is common amongst III-V semiconductors, with atoms tetrahedrally coordinated to atoms from the other chemical group.

The zinc-blende structure is similar to the diamond structure of bulk carbon (silicon, germanium) which also consists of two interpenetrating FCC lattices. The tetrahedral bond angle between the atoms is 109.5 degrees. The three classes of low index plane in the zinc-blende crystal are (111), (001) and (110), the latter being the growth surface we are concerned with. The bond length is

$$L = \frac{\sqrt{3}}{4} a_0, \tag{6.5}$$

where a_0 is the magnitude of the lattice constant.

The indium atom has a larger covalent radius than gallium and InAs has a greater lattice constant than GaAs. In table 6.1 we present the experimental values for the lattice parameters of the semiconductors and their corresponding bond lengths.

The misfit of two crystals is defined in terms of their individual lattice con-

Element	lattice constant Å	bond length Å
GaAs	5.65	2.45
InAs	6.06	2.62

Table 6.1: Experimental lattice constants and bond lengths of GaAs and InAs (Å).

stants;

$$\epsilon_0 = \frac{(a_{epilayer} - a_{substrate})}{a_{substrate}}. \tag{6.6}$$

According to this definition $\epsilon_0 = 7.3\%$ for InAs/GaAs and the strain is compressive.

6.3.2 Technical Convergences: Bulk Calculations

When performing a plane wave DFT calculation one should obtain energies and geometries acceptably converged with respect to the various parameters involved. If they are not comparisons of quantities such as surface energies between calculations might be unreliable. For example if we wish to compare the (110) surface energies of our two semiconductors we should use the same k point mesh and plane wave cut off in each separate computation. Ideally, these settings would be fully converged but limitations on cpu time make this an extravagant way of working. Thus even when the energies themselves are not absolutely converged cancellation of the absolute errors between calculations is relied upon to justify results. In DFT the error between calculations performed using the same parameter set often converges faster than absolute values within the calculations themselves.

Increasing the plane wave energy threshold will in general lower the energies obtained, allowing for numerical accuracy. Once the threshold reaches a

certain value the energy will be well converged and change very little with any further increase. This is also (necessarily) true of system geometries, and our tests confirm stability of the cohesive energy and theoretical lattice constant with respect to the plane wave cut-off. During the course of these calculations we will be considering small differences between the energies of large systems containing hundreds of atoms, and an error in the bulk energy per atom larger than a few meV will affect the validity of our results.

6.3.3 DFT Method

We used the VASP code [57] to perform our DFT calculations and also the included library of ultrasoft pseudopotentials for both GGA and LDA calculations. The GGA exchange correlation functional is the PW91. The plane wave cut offs for our calculations were set to 13.26 Ry to avoid aliasing errors on the FFT grids. This figure is derived by multiplying the maximum pseudopotential energy (As) by 1.3, as recommended in the VASP guide for high precision calculations of quantities like cleavage energies. Using this cut-off we obtained bulk cohesive energies and also surface energies in good agreement with the previous literature which we will discuss in more detail below. During geometry optimizations forces below a maxiumum of 0.01 eV/Å were considered converged.

When performing calculations on the III-V's we used the 3 electron pseudopotentials for Ga and In rather than the 13 electron pseudopotentials which also include the d electrons. Although the latter would have been preferable the large numbers of atoms in our simulations (up to 350) persuaded us to use the 3 electron pseudopotentials. We compare the bulk cohesive energies and

lattice constants of both pseudopotentials in subsection 6.3.6 and find them to be close in value.

6.3.4 K Point Convergences: GaAs

As well as an adequate plane-wave threshold one must also choose a suitable k point grid, which dictates the accuracy of numerical quadratures done in reciprocal / k space. In general the more k points we have the more accurate our integrations become, though convergence with respect to this parameter is not monotonic.

The fineness of the mesh required is dependent on supercell size, as distances in k space are the reciprocal of those in real space. Thus, whereas a single k point at the origin might be sufficient for the energy of a large supercell, one would need many more k points to converge the energy of a supercell the size of the primitive unit cell.

We use a cubic cell containing eight atoms for our calculations on the bulk semiconductors. To converge the k point sampling we compared energies gained with successively finer cubic meshes, each mesh having the same number of kpoints along each lattice direction, increasing from 111 (gamma point) to 222, 333 and so on.

The figures 6.2 and 6.3 show that for our eight atom cell an 888 cubic k point mesh is sufficient (we have calculated with the LDA and GGA). We have shown LDA and GGA results on separate graphs as the scale of the error is much smaller than the difference in absolute energy between the two approximations.

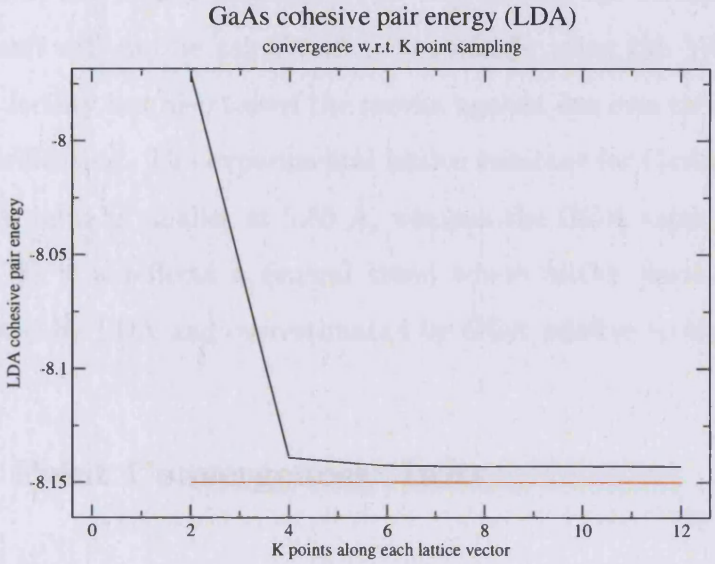

Figure 6.2: GaAs LDA cohesive energy (eV) w.r.t. k point sampling, red line - LDA cohesive pair energy obtained by Fuchs et al.[59]

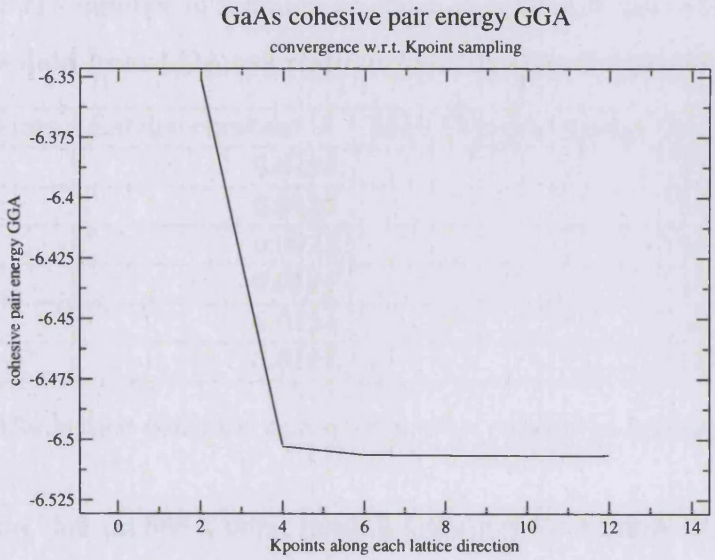

Figure 6.3: GaAs GGA cohesive energy (eV) w.r.t. k point sampling.

The (888) k point mesh ensures good convergence of the bulk lattice constant with an error of less than a milliangstrom. The equilibrium lattice constant for a cubic unit cell can be calculated automatically using the VASP code, we used this facility but also tested the results against our own strain energy curves for verification. The experimental lattice constant for GaAs is 5.65 Å and the LDA value is smaller at 5.59 Å, whereas the GGA value of 5.72 Å is larger. This is a reflects a general trend where lattice parameters are underestimated by LDA and overestimated by GGA relative to experiment.

6.3.5 K Point Convergences: InAs

We also performed tests on InAs in case a finer k point mesh is required for this material. For calculations involving both InAs and GaAs one would be obliged to use the more accurate mesh when they are combined into single simulation cell. The behaviour of both the lattice constant and cohesive energy versus the number of k points is shown in table 6.2 and table 6.3, the first contains data from LDA calculations and the second from GGA.

Kpoints	Lattice constant Å	LDA Cohesive energy (eV)
2	6.0193	-7.536
4	6.0120	-7.548
6	6.0123	-7.552
8	6.0124	-7.552
10	6.0124	-7.552
12	6.0124	-7.552

Table 6.2: InAs lattice constant and energy with respect to k points (LDA).

The tests show that an 888 k point mesh is sufficient for accurate description of InAs, conveniently the basic properties of GaAs were also well converged at this sampling. We bear in mind that for larger simulation cells than our

Kpoints	Lattice constant Å	GGA Cohesive energy (eV)
2	6.1567	-5.932
4	6.1660	-5.940
6	6.1663	-5.943
8	6.1664	-5.943
10	6.1664	-5.944
12	6.1664	-5.944

Table 6.3: InAs lattice constant and energy with respect to k points sampling (GGA).

cubic eight atom cell a coarser k point grid may be sufficient, but the 888 grid tested here gives us a good upper limit on the possible number of k points that would be required in each direction.

6.3.6 Pseudopotential Comparison

In table 6.4 we summarise the cohesive energies and lattice constants of InAs and GaAs treated using GGA with both 13 electron and 3 electron pseudopotentials. As far as the bulk is concerned the different pseudopotentials give very similar results. For calculations on larger systems we have therefore chosen to use the 3 electron pseudopotentials for expediency.

Property	InAs 15e	InAs 3e	GaAs 15e	GaAs 3e
a_0 Å	6.183	6.166	5.746	5.722
Cohesive Energy (eV)	-5.87	-5.94	-6.43	-6.51

Table 6.4: Cohesive energy and equilibrium lattice constants of the semiconductors with 13 e and 3 e GGA pseudopotentials.

6.3.7 GaAs bulk results

Here we will compare the results obtained with our converged parameters against those of the previous literature, [58],[59],[60], [61],[62],[63], [64],[65]. In figure 6.3 we have shown the LDA cohesive energy obtained by Fuchs et. al. in red [59] at -8.15 eV, close to our LDA value of -8.14 eV. Fuchs et al use a 50 Ry plane wave cut off and Troullier-Martins pseudopotentials [50]. Juan et al [60] find an equivalent energy of -8.58 eV using LDA with an 18 Ry cut off and 4 special k points. These values are calculated relative to the spin polarised atomic ground state, the magnitude of the correction (as opposed to unpolarised ground states) being 1.59 eV per GaAs pair. The experimental value for the cohesive energy of GaAs is -6.52 eV [52], which is higher than the value given by LDA and the LDA lattice constant is found to be 5.59 Å, which is less than the experimental figure of 5.65 Å.

The cohesive pair energy obtained using GGA is less than that found with LDA and is -6.51 eV as opposed to -8.15 eV. This is in much better agreement with the experimental value of -6.52 eV [52]. Previous GGA calculations of the GaAs cohesive energy are also in good agreement with ours, Fuchs et. al. finding -6.63 eV while Juan. et. al. [60] obtain -6.51 eV. In their calculations Fuchs. et. al. also use an 888 kpoint mesh with a higher plane wave cut off energy of 50 Ry. The results for the cohesive energy of GaAs are summarised in table 6.5

The lattice constant obtained using GGA is 5.72 Å which is larger than both the LDA value (5.59 Å) and also experiment (5.65 Å). This reflects a general trend in DFT whereby lattice parameters are systematically underestimated by the LDA approximation and overestimated by GGA, whilst cohesive en-

approximation used	cohesive pair energy (eV)
LDA - Ours	-8.14
LDA - Fuchs	-8.15
LDA - Juan et.al.	-8.58
GGA - Ours	-6.51
GGA - Fuchs	-6.63
GGA - Juan	-6.51
Experiment	-6.51

Table 6.5: GaAs cohesive energies (LDA and GGA).

ergies tend to be overestimated by the LDA approximation.

The GGA energy agrees well with the experimental value, but the LDA energy is about 2 eV lower. Later we will look at more complex supercells in order to model dislocation cores. These calculations are too expensive for us to perform using both LDA and GGA, hence we choose to model the systems using only GGA which gives better cohesive energies than LDA, although we will see in sections 6.4.6, 6.4.7 that LDA provides more accurate surface energy estimates.

6.3.8 InAs Bulk Results

We repeat the calculations done for GaAs, finding the equilibrium lattice constants, the cohesive energy and also the (110) surface energy and geometry. The experimental value of the cohesive energy is -6.06 eV, with LDA we obtain a value of -7.55 eV and GGA gives us a higher value of -5.94 eV, which is closer to experiment. These energies are all relative to a spin polarised reference ground state.

The experimental lattice constant is 6.06 Å and we obtain a values of 6.01 Å

with LDA and 6.17 Å with GGA. Again GGA overestimates the lattice constant whilst producing an accurate cohesive energy. LDA overestimates the lattice constant but overestimates the cohesive energy by 1.5 eV (taking the minus sign into account).

6.3.9 InAs Under Biaxial Strain

Below we plot the energy curve of bulk InAs with the GaAs lattice parameters in the (110) plane, but with varying (110) interplanar spacing, figure 6.4 The

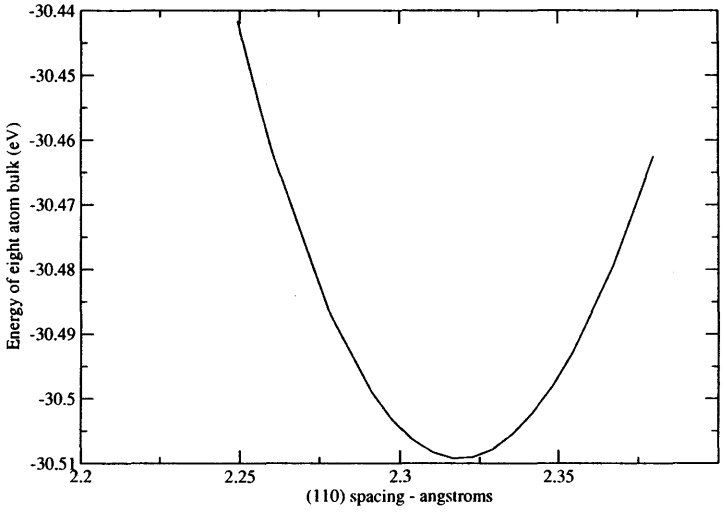

Figure 6.4: Energy Vs (110) interlayer spacing of biaxially strained InAs (GGA).

curve in figure 6.4 was produced using GGA (PW91), giving more accurate absolute cohesive energies than LDA. In the two directions of the plane, [001] and [1$\bar{1}$0], the InAs was constrained to the theoretical GaAs lattice parameter of 5.7216 Å, its own theoretical lattice parameter being 6.1664 Å. We find an equilibrium (110) spacing of 2.32 Å, which is 0.14 Å greater than that of InAs free from strain. The corresponding cohesive energy per pair of -7.63

eV is 0.19 eV higher (less stable) than that of unstrained InAs.

III-V crystal	(110) spacing Å	energy per pair
GaAs	2.02	-8.0
InAs	2.18	-7.82
Uniaxially strained InAs	2.27	-7.77
Biaxially strained InAs	2.32	-7.63

Table 6.6: (110) interlayer spacing of InAs under different strain conditions.

Although the strained InAs is placed initially at the GaAs lattice sites, the atoms undergo a small displacement away from these under the compressive strain. Thus the bond length in the (110) plane is no longer equal to that of bulk GaAs (2.48 Å), but increases slightly to 2.58 Å as a result of the displacements.

6.3.10 InAs Under Uniaxial Strain

Here we look at InAs constrained to the GaAs lattice constant in the [001] direction, and to a quasi-InAs lattice constant along $[1\bar{1}0]$, but free to relax along [110]. The lattice constant along $[1\bar{1}0]$ is not exactly that of GaAs as it is chosen from the structure of the simulation cells we use to study the edge dislocation network. We form the dislocations by removing a single row of InAs for every fifteen rows of GaAs in the supercell in accordance with estimates of the ideal dislocation spacing. As the dislocation network is one dimensional we do not have to consider misfit dislocations with line vector along the $[1\bar{1}0]$ direction as well. As the lattice constant of GaAs is 5.72 Å, the InAs is set at a spacing of 15/14 that of GaAs along $[1\bar{1}0]$, close but not exactly equal to the spacing in unstrained InAs.

Below we present a relaxation energy curve for the uniaxially strained InAs, from which was can derive the (110) interlayer spacing, and the corresponding energy per InAs pair in (figure 6.5 - note that energies shown are without spin correction).

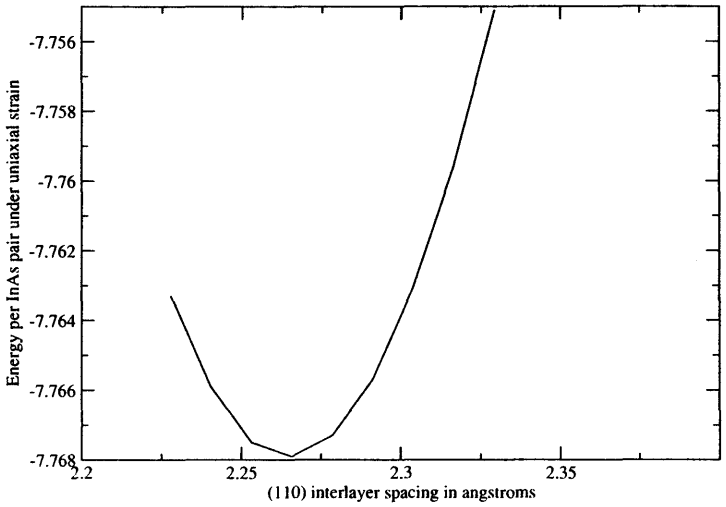

Figure 6.5: Energy w.r.t. (110) interlayer spacing of uniaxially strained InAs (GGA).

We see that the minimum energy, -7.77 eV, lies between that of relaxed bulk InAs (-7.82 eV) and biaxially strained InAs (-7.63 eV). Although the edge dislocations relieve strain along $[1\bar{1}0]$ there will be considerable residual strain energy left in the InAs/GaAs bicrystal even after formation of the edge dislocation network. Although this strain is eventually relieved by the onset of a network of 60 degree dislocations they do not appear until many tens of epilayers have been deposited, and we do not consider this aspect of the strained heteroepitaxy.

The equilibrium interplanar spacing, 2.27 Å, is also between that of the relaxed bulk (2.18 Å) and when biaxially strained (2.32 Å). This is what we would expect, since the degree of vertical expansion correlates with the total

amount of strain. Later on we will calculate the geometries of InAs/GaAs bicrystals containing edge dislocations, and the figures we have obtained will be useful in analysing our results, for example to compare InAs interlayer spacings after dislocation formation against the ideal values already derived, in table 6.7.

InAs	(110) spacing (Å)	energy per pair (eV)
unstrained	2.18	-7.82
uniaxial strain	2.27	-7.77
biaxial strain	2.32	-7.63

Table 6.7: Equilibrium (110) spacings and energies of InAs under different strains

6.4 Properties of the (110) Surface

The (110) surface is the growth surface of the strained bicrystal which we will examine. It contains equal numbers of group III and group V atoms, and is charge neutral overall, being a non-polar surface. When it is cleaved the valency of the group V atoms can be satisfied by bonds to the group III atoms, and the surface relaxes rather than reconstructing. The periodicity of the relaxed surface is the same as that of the underlying bulk. This is in contrast with, for example, the (001) surface. When the zinc blende lattice is cleaved to reveal the (001) surface it contains atoms of either group III or group V but not both. The (001) surface is non-stoichiometric as there are different numbers of each species present compared to the bulk. As a consequence of this the (001) surface displays a more complex set of reconstructions than (110) as the atoms attempt to satisfy their bonding

requirements, and the resulting reconstructions are of lower periodicity than to the underlying bulk. For this reason (110) planes are often chosen as cleavage planes of III-V semiconductors due to their more stable behaviour.

Experimental observations [62],[65] reveal that the group V atoms of the (110) surface are pushed upwards out of the surface plane whilst the group III atoms are drawn inwards, showing up as a characteristic tilt of the surface unit cell as in figure 6.6. The strain effects of this tilt do not penetrate more than a few layers into the underlying bulk, unlike the reconstruction effects of more complex surfaces.

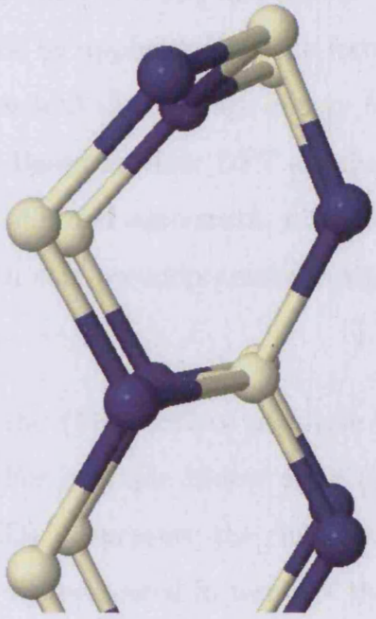

Figure 6.6: InAs (110) surface, blue atoms are Indium, white Arsenic

6.4.1 Previous Literature on GaAs and InAs (110) Surfaces

Much previous work has been done on GaAs which is seen as an archetypal system for the study of III-V semiconductor properties. Qian et al [61] completed a study of the (110) surface energy and geometry using DFT (LDA) performing the first calculation of the cleavage energy of a heteropolar semiconductor surface using density functional theory. An experimental meaasurement of the GaAs(110) surface energy is provided by Messmer and Bilello ([64]) who perform a fracture experiment on the GaAs crystal.

Using a spark discharge method Messmer and Billelo fracture the crystal and expand the resulting crack by applying a tensile force perpendicular to the cleavage plane. They measured the cleavage energy of the GaAs(110) plane to be 0.86 J/m^2 (± 0.15). Based on their DFT approach Qian et al obtained an energy of 0.91 J/m^2, in good agreement with the experimental result. Qian et al used LDA with soft pseudopotentials and a 6 Ry energy cut-off during their calculation.

The actual geometry of the (110) surface of GaAs has been presented in numerous publications. For example Meyer et al ([62]) apply low energy electron diffraction (LEED) to measure the characteristic displacements of the surface. The geometry is presented in terms of the displacements shown in figure 6.7 which shows a diagram of the surface (this figure is based on that of Meyer et al)

More recently Alves et al ([58]) have published results on the atomic structures and electronic properties of III-V (110) semiconductor surfaces using

DFT (LDA). They focus on GaP, InP, GaAs and InAs for their calculations, using a cut-off of 18 Ry, with 4 special k points. The GaAs (110) surface parameters calculated by [61] and [58] are tabulated below in table 6.8 as well as those from experiment [62]. The different sets of data are in good agreement, though the calculations systematically underestimate the characteristic parameters and later in this chapter we will compare the results of our own calculations on the (110) surface against these.

author	d1p	d1x	d2p	d12p	d12x
Qian et. al.	0.58	4.390	0.07	1.440	3.180
Alves et. al.	0.67	4.407	0.098	1.415	3.180
Meyer et. al. (expt)	0.69	4.518	0.120	1.442	3.339

Table 6.8: Characteristic geometric parameters of the GaAs (110) surface as in diagram 6.7

In [58] the authors also examine the characteristics of the (110) surface of InAs. Though they do not calculate the cleavage energy they do obtain the geometric parameters through density functional LDA calculations, and they compare these parameters against those obtained using LEED ([65]), their results are shown in table 6.9 below.

author	d1p	d1x	d2p	d12p	d12x
Alves et. al.	0.75	4.663	0.128	1.445	3.395
Mailhiot et. al. (exp)	0.78	4.985	0.140	1.497	3.597

Table 6.9: Characteristic geometric parameters of the InAs (110) surface.

Previous density functional calculations on the InAs (110) surface subject to strain have also been performed by Moll et. al. [63], in relation to equilibrium shapes of InAs quantum dots. In [63] the authors calculate the cleavage energy of the InAs (110) surface using DFT (LDA) and also calculate the components of the surface stress tensor along the two directions of the (110)

plane, [001] and [1$\bar{1}$0], denoting them σ_x and σ_y respectively. These components of the surface stress enter into 6.7. We summarise the results of [63] in table 6.10 below. The positive sign of the stress tensor components indicates that the InAs (110) surface energy should lower under compressive strain.

$$\gamma^{strained} = \gamma + \Sigma_{ij}\sigma_{ij}\epsilon_{ij} + Higher\ Order\ Terms \tag{6.7}$$

InAs(110)	γ	σ_x	σ_y
Moll et. al.	41	26	54

Table 6.10: InAs LDA cleavage energy (γ) as calculated by Moll et al (meV/$\AA^2$).

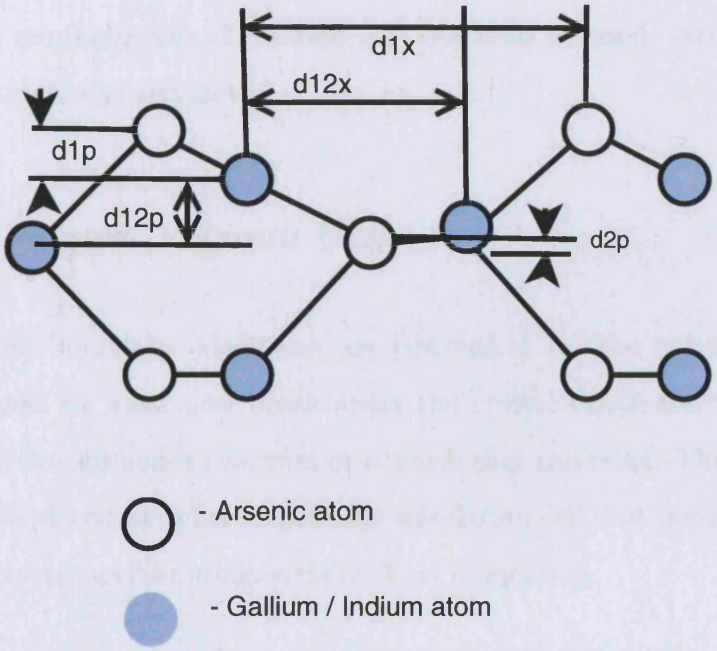

Figure 6.7: Characteristic displacements of III-V (110) surface.

6.4.2 Technical Convergences: Surface Calculations

A good understanding of the (110) surface behaviour of both materials will also be crucial in our later studies, and conditions necessary for accurate simulation of the surface are discussed here. To reproduce the surface geometry sufficient underlying bulk material must be included in our calculations to capture the accommodation of the surface strain field into the substrate. As we are using periodic boundary conditions we cannot simply place a semi infinite vacuum above the surface either, and instead use the repeating slab approximation, where a large enough vacuum gap effectively zeroes the interaction between neighbouring slabs. Below we establish the amount of material underlying the surface which we must include in our calculations and also the minimum size of vacuum gap that can be used, verifying the details for both GaAs and InAs.

6.4.3 Minimum Vacuum Gap

While periodic boundary conditions are convenient for the calculation of bulk properties, we must now break apart the crystal inside the repeating cell, exposing two surfaces to vacuum into which they can relax. The vacuum region may be placed anywhere inside the simulation cell, but must be large enough to prevent neighbouring surfaces from interacting.

Increasing the vacuum gap also increases the size of the supercell, and calculations become more expensive due to the additional plane waves required. Therefore we seek the minimum gap sufficient to negate interaction between repeating slabs. In figure 6.8 below we show the energy of a 12 atom (GaAs)

surface slab as the gap between slabs is increased up to 8 Å. Note that the interlayer spacing is 2 Å and that this is the initial separation distance of the two surfaces. Once the gap is larger than 5 Å there is little change in the total energy (roughly a hundredth of an eV per atom), so we maintain a minimum gap of 5 Å between neighbouring surfaces in all our calculations (the energies shown are for a twelve atom GaAs surface slab using GGA).

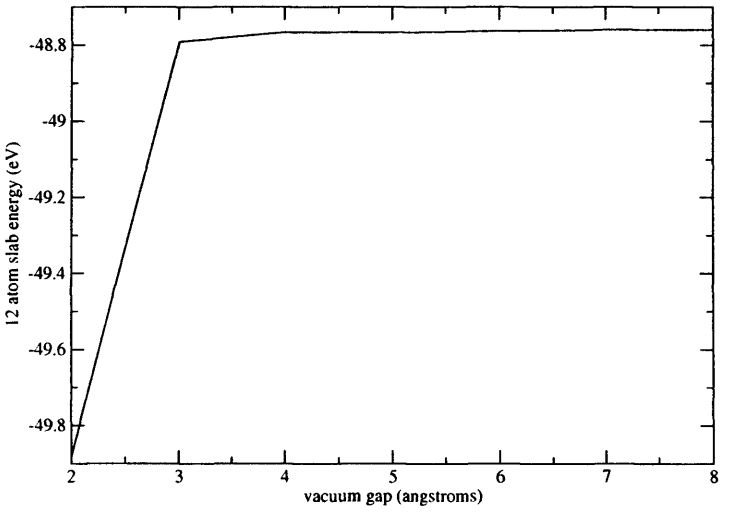

Figure 6.8: Total energy (eV) of 12 atom surface slab Vs vacuum separation.

With a sufficiently large gap the total energy should be independent of the number of k points which we use to sample along the surface normal. The total energy calculated with a k point mesh of 8x8x8 should be identical to that calculated with a mesh of 8x8x1, and we confirmed this to be the case.

6.4.4 Number of Layers in the Surface Slab

In the finite slab approximation we must also be concerned with slab depth. As the surface reconstructs or relaxes it creates a strain field which penetrates into the layers underneath. If there are insufficient layers beneath the surface

of the finite slab the action of the bulk in accomodating the surface strain may be poorly approximated, and the surface geometry will be inaccurate as a result.

When placing a vacuum gap inside the supercell we create two exposed surfaces. It is not strictly necessary to relax each of these, for example one may choose to terminate one surface with hydrogen, freezing the atoms in bulk-like positions, and then relax the geometry of the remaining free surface. We however have chosen to relax both surfaces, as in this way we can be sure that there is sufficient bulk crystal by monitoring the relaxation on each side, checking that the geometries are symmetric, and that the central portion of the slab behaves like bulk crystal.

The two surfaces should be geometrically identical once the slab is thick enough, and the slab itself should be mirror symmetric about the centre. Our tests show that a slab having seven atomic layers in total (i.e. five bulk like layers in between the two surface layers) is sufficient to converge both surface geometries to within an acceptable tolerance of roughly 1% in each parameter. We monitor several interatomic distances to ensure convergence of the calculated geometries, shown in figure 6.7 the values of the parameters themselves are reported in table 6.11 as a function of the total number of layers.

Another important test of our finite slab approximation is the energetic cost of additional layers. This should be equal to the energy of an equivalent amount of bulk crystal, proving the material in the slab centre to have bulk properties, and that the slab is sufficiently thick for surface strain not to distort its central layers. In figure 6.12 we show the energy difference between surface slabs with increasing numbers of layers. That is, the figure shown at

N layers	d1p	d1x	d2p	d12p	d12x
5	0.677	3.808	0.079	1.479	3.268
6	0.697	4.474	0.130	1.444	3.254
7	0.685	4.470	0.104	1.463	3.259
8	0.692	4.472	0.116	1.454	3.258
9	0.688	4.471	0.111	1.458	3.258
10	0.690	4.471	0.114	1.457	3.258

Table 6.11: GaAs surface geometric parameters w.r.t. slab thickness (Å).

10 layers is the difference in the total energy of a slab containing 10 layers altogether and a slab which contains 9 layers. We see from table 6.12 that the energies do eventually converge towards the equivalent bulk value. An 881 k point was mesh used to calculate the energy differences shown in the table.

In figure 6.9 we plot the slab energy difference as a function of the number of layers for up to twenty layers, the convergence can be clearly seen after ten layers. As this calculation was a proof of principle we used a cheaper 661 k point mesh because we are not concerned with the absolute energy values, but instead with demonstrating the convergence of energy differences.

N layers	Bulk Energy	Slab Energy	$\Delta(slab - bulk)$	$\Delta(bulk)$	$\Delta(slab)$
6	-50.484	-48.775	1.71	na	na
8	-67.312	-65.600	1.71	-16.828	-16.824
10	-84.140	-82.426	1.71	-16.828	-16.827
12	-100.968	-99.255	1.71	-16.828	-16.829
14	-117.796	-116.081	1.72	-16.828	-16.826
16	-134.624	-132.910	1.71	-16.828	-16.828

Table 6.12: GGA energies (eV) for GaAs bulk and slabs, one atomic pair per layer (881 kpts).

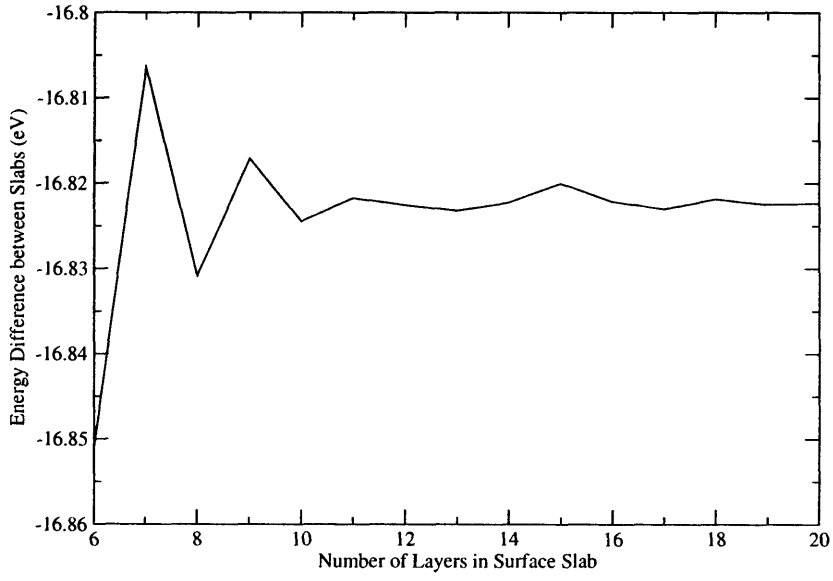

Figure 6.9: GaAs slab energy differences towards w.r.t. slab thickness. (2 GaAs pairs per layer, using a 661 k point mesh). The equivalent energy of a GaAs from an eight-atom bulk cell with the same k point mesh is -16.825 eV.

We repeat these tests for InAs (110), again checking the slab thickness necessary to converge the surface energy and geometry. We tabulate the characteristic distances (table 6.13) and also plot the energy differences between successive slabs in figure 6.10.

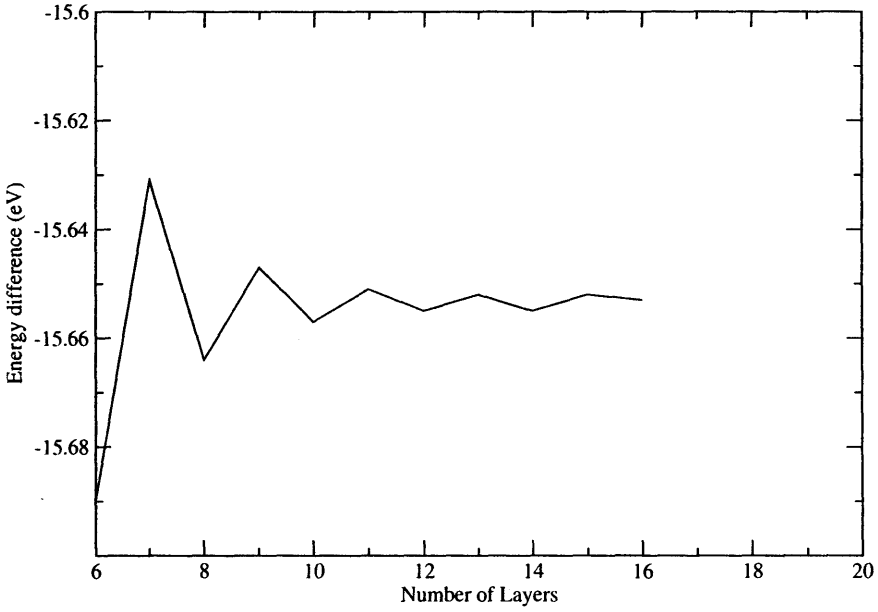

Figure 6.10: Convergence of InAs slab energy differences towards the relaxed bulk value (GGA).

N layers	d1p	d1x	d2p	d12p	d12x
5	0.670	4.454	0.085	1.483	3.264
6	0.703	4.472	0.160	1.426	3.243
7	0.683	4.464	0.121	1.457	3.251
8	0.694	4.466	0.142	1.440	3.248
9	0.690	4.466	0.130	1.450	3.250

Table 6.13: Geometric parameters (Å) for InAs (110) surface (see figure 6.7).

6.4.5 Convergence Test Summary

We have established that a plane wave threshold of 13.26 Ry and an 888 k point mesh are sufficient for modelling the bulk solid with an eight atom cell. Tests have been done to demonstrate the convergence of the cohesive energy and lattice constant for both InAs and GaAs with respect to these parameters.

For surface slab calculations a vacuum gap of three angstroms prevents unwanted interactions between repeating slab surfaces, and the energies calculated with an 881 and an 888 k point mesh are equal with this vacuum gap. In order for the central layers to behave equivalently to bulk crystal we must include no fewer than seven layers in total, this has been shown by monitoring the surface geometry and also that the cost of adding additional layers to the slab converges towards the energy of the same amount of bulk crystal. The slab energy differences are not fully converged by seven layers with an error of about 0.01 eV per atom relative to the converged result, but due to the limits on available cpu time when modelling the InAs/GaAs(110) bicrystal we include seven layers of GaAs in the calculations to reproduce the action of the bulk substrate, though inclusion of further layers would have been desirable. For example in figure 6.9 we see that the central layers of the slab have a bulk-like energies only from eleven layers onwards, before which point there are deviations of up to 0.005 eV per GaAs pair from the converged energy difference.

6.4.6 GaAs (110) Surface Energy

The energy of a slab with two surfaces can be written as follows,

$$E_{slab} = E_{bulk} + 2E_{surface}, \tag{6.8}$$

the slab may be viewed as a section of bulk crystal with an additional (positive) energy due to relaxation of its two surfaces. To isolate $E_{surface}$ we can obtain E_{bulk} in a separate calculation and subtract it from E_{slab} [61]. A theoretically equivalent approach would be to set the bulk energy per layer equal to the converged slab energy difference and then subtract this from the total slab energy, leaving the energy of the two slab surfaces as the remainder.

We calculate the surface energy of GaAs using the first technique, where the energy of the bulk crystal is obtained in a separate calculation. Using LDA data we find a surface energy of 50 meV/Å^2. This is in good agreement with previous calculations which also used LDA, for example Moll et al. [63] find a value of 52 meV/Å^2, and Qian et al. [61] find 57 meV/Å^2. The experimental value, found through fracture experiments is 54 $\pm$9 meV/Å^2 [64]. Using GGA we find a lower value of 38 meV/Å^2, so the GGA approximation tends to underestimate the (110) surface energy. However we anticipate that the absolute errors in the surface energies will not affect the results of our dislocation core studies due to cancellation of absolute errors in comparing different simulation cells.

6.4.7 InAs (110) Surface Energy

We estimate the energy of InAs (110) using the same method employed for GaAs (110), taking advantage of equation 6.8 to isolate the surface energy. Within the GGA approximation we find an energy of 32.1 meV/Å^2 which is lower than the (110) surface energy of GaAs (38 meV/Å^2). Within the LDA approximation we find a value of 41.2 meV/Å^2, which is in good agreement with that of Moll et. al., at 41 meV/Å^2. As for GaAs we find GGA to give a lower value for the surface energy than LDA. Below we show LDA and GGA (110) surface energies that we have calculated for InP and GaP as well as InAs and GaAs (showing that the trend is consistent across a range of different III-V semiconductors) table 6.14.

III-V	LDA surface energy (meV/Å^2)	GGA surface energy (meV/Å^2)
InP	50.2	37.9
GaP	59.2	48.5
InAs	41.2	32.1
GaAs	50.0	38.1

Table 6.14: LDA, GGA (110) surface energies for a range of III-V semiconductors.

6.4.8 Biaxially Strained InAs (110) Surface Energy.

We also evaluate the surface energy of InAs (110) under biaxial strain, using the method already applied to GaAs and unstrained InAs. We construct the simulation cell by placing the InAs into a surface slab with GaAs lattice positions and then relax the InAs. We must be careful to include enough vacuum to allow for the vertical expansion of the InAs. For example if each layer expands upwards by 0.2 Å, and we have eleven layers of InAs in our

slab there will be a total expansion of 2.2 Å. If the initial vacuum gap is just three Å, the vertical expansion will erode it down to 0.8 Å, which is too small (see figure 6.8). Thus we set a vacuum gap of at least 6 Å in our calculations of strained InAs properties, after relaxations were completed it was ensured that there were at least 3 Å of vacuum remaining.

We show a graph (figure 6.11 plotting energy differences between biaxially strained InAs slabs with increasing numbers of layers. The energy difference converges towards the calculated cohesive energy of the biaxially strained bulk, which was -7.63 eV. The value of the surface energy is 25 meV/Å^2. This

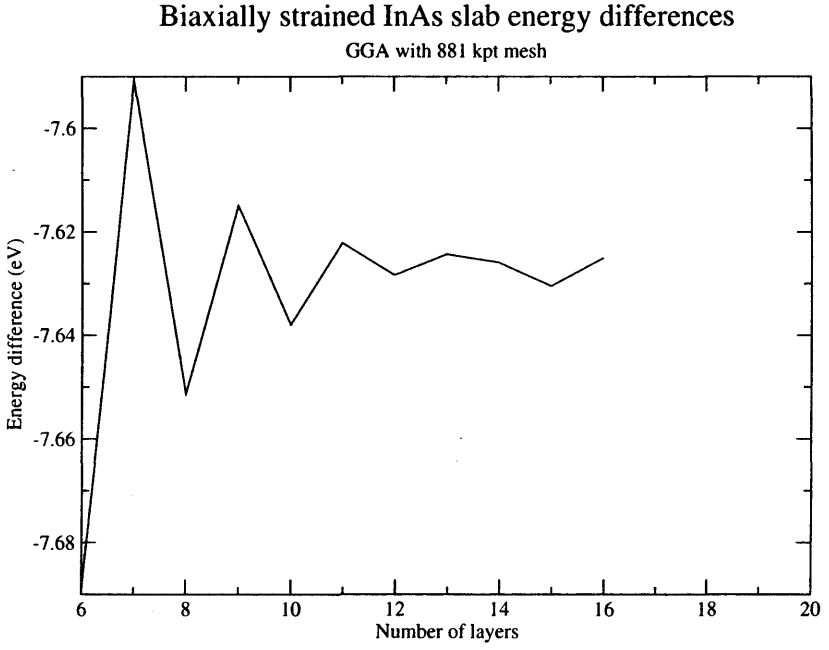

Figure 6.11: Convergence of slab energy differences for biaxially strained InAs (eV).

is lower than the unstrained InAs (110) surface energy of 32 meV/Å^2. The strained surface energy was calculated using the formula 6.8 but substituting E_{bulk} with an appropriate multiple of the slab energy difference. Straining the InAs leads to a lowering of the (110) surface energy, by 7 meV/Å^2. The

surface energy of strained InAs(110) has previously been examined by Moll et. al. ([63]). There they write an equation for the energy of the strained surface relative to the area of the undeformed surface,

$$\gamma^{strained} = \gamma + \sum_{ij} \sigma_{ij}\epsilon_{ij} + h.o.t. \tag{6.9}$$

(where h.o.t. are higher order terms) to first order in the stress and strain tensors. Using the LDA approximation they derive a contribution from the first order correction of 6 meV/Å^2 for the InAs(110) surface strained at GaAs lattice constants. They find the components of the surface stress tensor to be tensile, specifically that σ_x = 26 meV/Å^2 and that σ_x = 54 meV/Å^2, since compressive strain has a negative sign this implies a lowering of the (110) surface energy, in agreement with our observations.

Another effect of the compressive strain is to increase the number of slab layers required to converge the slab energy differences to twelve as opposed to eight for the slabs with the theoretical equilibrium geometry. The compressive strain exaggerates the slight tilting of the central layer which manifests as oscillatory energy differences when the number of layers in the slab are too few.

6.5 Edge Dislocation Calculations

There are two main points which we wish to establish through our calculations. Firstly, to identify the core structure of the edge dislocations which form a network in the early stage of InAs/GaAs(110) heterepitaxy. We will do this by comparing the energies of different candidate structures and seeing

which is lowest. Belk et al suggest that the dislocations must form in the layer immediately beneath the surface and must thus lie in the second layer or even above [16], whereas Oyama et al assume from their experiments and calculations that the dislocation cores should lie in the first InAs epilayer. Density functional theory does not give the full picture of the heteroepitaxial growth process but we can establish which layer will be most stable for the dislocation core to lie in. Having established which layer is most stable we can then turn to the question of the dislocation symmetry plane. The zinc-blende structure involves alternating In and As atoms, and the dislocation core may be formed either over In or As, whichever is energetically most stable. Finally having established the lowest energy core configuration we will be able to obtain the critical epilayer thickness at which the dislocation network lies lower in energy than the equivalent amount of coherently strained InAs. As the dislocations are formed by removing rows of InAs pairs from the covering epilayers we will have to compensate for the missing InAs pairs by adding in a "chemical potential" term to the energy of the supercells containing the edge dislocations.

6.5.1 Equilibrium Misfit Dislocation Spacing

In order to construct the simulation cells containing edge dislocations we must first estimate the equilibrium spacing between dislocations. Experimentally the dislocations form an array along the [001] direction, separated by an average of 60 Å along $[1\bar{1}0]$ as in figure 6.1. We can predict the separation by considering the lattice mismatch and removal of the InAs row which is required for their formation from the bulk crystal. The width of the removed row must be the lowest integer multiple of the mismatch size

(this is a condition of commensuration). We must remove an InAs pair in order to form the dislocation, and that this will leave a corresponding gap of $W = 4.04\ \mathring{A}$ (using the experimental InAs lattice constant of 6.06 $\mathring{A}$). The following relation must then be satisfied,

$$W = Nm \tag{6.10}$$

where m is the lattice mismatch (angstroms) between InAs and GaAs in the $[1\bar{1}0]$ direction and N is an integer equal to the number of GaAs rows. The mismatch along $[1\bar{1}0]$ is (for the zinc blende structure)

$$\begin{aligned} m &= (L_{InAs} - L_{GaAs}) * \frac{2}{3} \\ &= (6.06 - 5.65) * \frac{2}{3} \\ &= 0.273\mathring{A}. \end{aligned} \tag{6.11}$$

Substituting m into equation 6.10 gives us $N = 15$ (rounded to the nearest integer), so we expect an edge dislocation to form once in every fifteen rows of GaAs (on average). As the GaAs row width is 3.77 Å, this corresponds to an equilibrium edge dislocation spacing of $d = 56.5$ angstroms, which corresponds well with the experimental value (about sixty angstroms).

An alternative way to derive d is in terms of the fractional lattice mismatch and the Burgers vector of the dislocation, using the following equation;

$$d = \frac{b}{\epsilon_0} \tag{6.12}$$

where b is the magnitude of the Burgers vector. For an ideal edge dislocation with Burgers vector $B = (a_0/2)[1\bar{1}0]$, $b = a_0\sqrt{2}$, and for InAs b = 4.285 Å.

The value of ϵ_0 is 0.0726, so that the predicted d becomes 59.1 Å. Again this agrees closely with the experimental value. There is a small discrepancy in the two solutions of 2.6 Å, this results from our rounding to the nearest integer the number of rows of GaAs at which edge dislocations may form (fifteen), the number of rows is a discrete quantity, not fractional.

However, the GGA lattice constants of the materials differ from the experimental values, and consequently the degree of strain in our calculations will also be different. The GGA lattice constants are slightly larger than the experimental values, and the theoretical strain is $\epsilon_{th} = 0.0786$. Using the formula 6.12 we can estimate the equilibrium dislocation spacing suggested by the GGA lattice constants, obtaining $d_{gga} = 55.46$ Å, with the magnitude of the Burgers vector $|b_{th}| = 4.36\mathring{A}$. This is slightly less than the experimental equilibrium dislocation spacing because the theoretical value of the mismatch strain, $\epsilon_{th} = 7.9\%$, is higher than the experimental value of 7.2%. The LDA strain is closer to the experimental value at 7.5% but we perform the dislocation calculations using GGA in order to obtain cohesive energies which are closer to the experimental values.

For our calculations we choose a periodic supercell with an edge dislocation occurring every fifteen GaAs pairs, with our altered lattice constants this corresponds to a spacing of d = 57.2 Å, which is close but not equal to the experimental spacing of 60 Å.

6.5.2 Increased Supercell Size

We have seen that the ideal spacing between edge dislocations in strained InAs/GaAs(110) is 60 Å, corresponding to a width of fifteen GaAs pairs. To

InAs epilayers	wide energy	wide energy/row	small energy
2	-1090.8009	-72.720	-72.728
3	-1205.1566	-80.344	-80.354
4	-1319.9400	-87.996	-88.006
5	-1434.0182	-95.601	-95.622

Table 6.15: Energies (eV) of surface slabs with fifteen III-V rows and only a single row wide.

model the system we will need a supercell of this width, with sufficient layers to accurately model the chemistry of the heterointerface. In our previous calculations surface slabs have only been as wide as a III-V pair in $[1\bar{1}0]$, we have been able to use this approximation because of periodic boundary conditions, with eight kpoints along the $[1\bar{1}0]$ direction. The fineness of the k point grid required to converge the energy per atom decreases with the real space size of the supercell along a given axis. For example, when modelling surface geometries we saw that only a single k point was needed along the surface normal to converge the supercell energy, and that adding more kpoints produced no effect. We must check that the wider supercells reproduce the characteristics observed in our narrow surface cells before we perform more complicated calculations with them. Table 6.15 shows the total energy for wide (fifteen GaAs rows) and narrow (a single GaAs row) supercells (all energies are in eV and do not include corrections for the spin polarised atomic ground states - however such corrections will cancel out when we are comparing total energies). The results in 6.15 show that the energy per III-V row is in good agreement between our fifteen row cells with an 811 kpoint mesh and our single row cells with a 881 kpoint mesh. In addition the energy of strained bulk InAs extrapolated from the wide cells comes out as -7.63 eV (without spin correction), which agrees exactly with that obtained using the smaller supercells.

6.5.3 Coherent Epilayer Growth

Having already studied the properties of InAs subject to biaxial strain in a previous section, we have now calculated the relaxed energy and geometry of a slab containing seven layers of GaAs substrate and from one to five layers of InAs on top. One effect of the compressive strain is to increase the (110) interlayer spacing of the InAs above its relaxed bulk value, we can see this effect in figure 6.13 where the InAs epilayers (blue and white) clearly have a greater interlayer spacing than the 2.18 Å of relaxed InAs. We also see from the figure that the InAs surface dimer bonds are shorter than those of the GaAs surface underneath. We previously established that the effect of compressive strain was to lower the (110) surface energy of InAs, as the shortened InAs dimer bonds lead to increased energetic stability at the cost of increased strain energy in the underlying bulk.

We re-confirm that 8 kpoints are sufficient along [001] for our new 15 row wide supercell, which we will use to model the misfit dislocations, checking that the energy is well converged. From the graph 6.12 of the total energy from 2 to 16 kpoints we see that the variation is still quite large until we have 6 kpoints or more along [001]. The plot below confirms that the energy of the wide cell is well converged with an 811 kpoint mesh, as we would expect from our calculations on the bulk.

6.5.4 Vertical Expansion

In figure 6.13 we show some measurements of the (110) interlayer spacing. There is an increase of about 0.25 Å moving from the InAs-GaAs interface

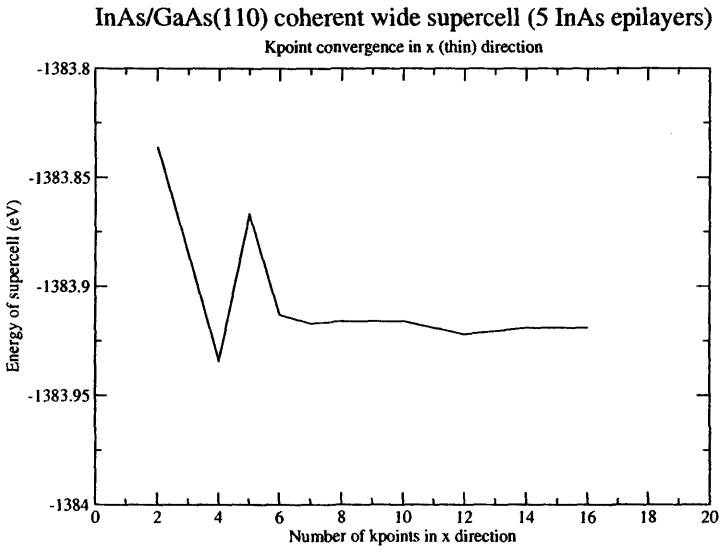

Figure 6.12: Total energy Vs k points for (15 row) wide supercells.

to the pure InAs epilayers. We may estimate an average layer spacing of 2.3 Å from the measurements in the diagram, which agrees well with our previous result of 2.32 Å being the equilibrium interlayer spacing for InAs under equivalent biaxial strain. The small difference/tilting in InAs spacings shown on the diagram may be ascribed to the strain effects of the tilted InAs surface dimers.

6.5.5 Increasing Strain Energy

We want to compare the energy of the coherently grown InAs against the semi-coherent InAs containing a dislocation network as a function of the total epilayer number. As a first step we calculate the energy of InAs lying coherently on the GaAs substrate, the results are shown below in table 6.16; The residual strain energy of the slabs is calculated by subtracting away the energy of the equivalent amount of relaxed bulk material, the values

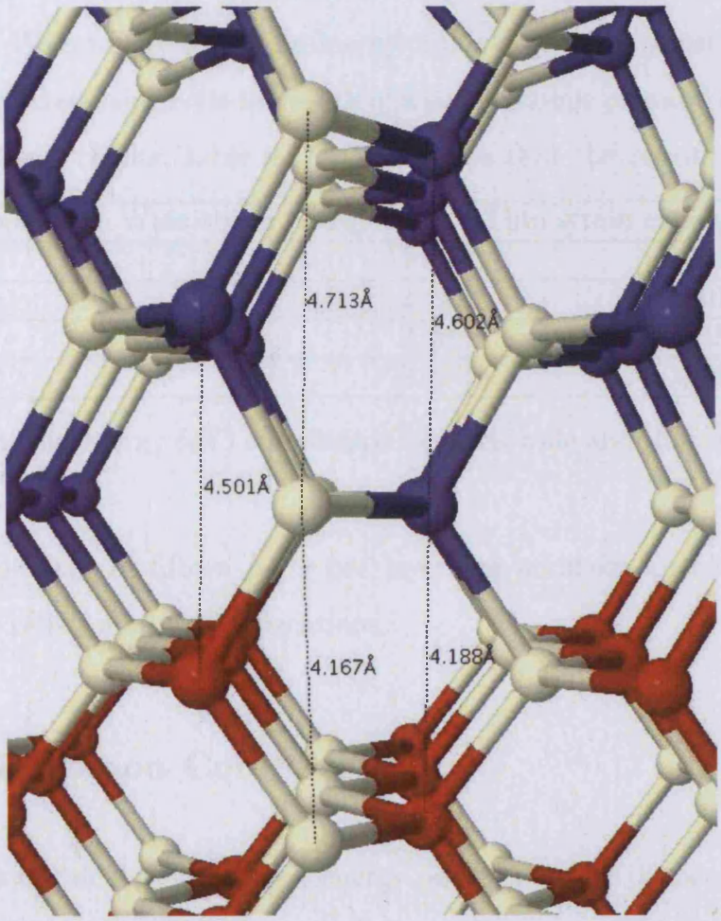

Figure 6.13: Interlayer spacings of compressed InAs shown above

n epilayers	total energy	strain energy
2	-1090.801	27.31
3	-1205.157	30.33
4	-1319.940	32.93
5	-1434.018	36.22

Table 6.16: Energy (eV) of coherent InAs epilayers on GaAs substrate (15 pairs per layer).

(without spin correction) are -7.82 eV per InAs pair and -8.41 eV per GaAs pair. Note that the total strain energies shown include the surface relaxation energies too. We compare the strain energy of the large cells against the same quantity calculated using cells the width of a single atomic pair and find good agreement of our results, table 6.17. This shows that the results obtained

n epilayers	Wide strain energy /(15)	Thin strain energy
2	1.82	1.81
3	2.02	2.01
4	2.20	2.19
5	2.41	2.39

Table 6.17: Strain energy (eV) comparison between wide and thin cells (InAs on GaAs).

with the wide cells of fifteen pairs per layer are accurate for comparison against wide cells containing dislocations.

6.5.6 Dislocation Core Geometry

Now we wish to calculate the lowest energy position for the dislocation core. This may be decomposed into two questions, firstly in which layer is the dislocation most likely to sit? Belk et.al. [16] suggest that the core should sit in the second layer or above. However Oyama et.al. [53] suggest that it lies in the first layer at the heterointerface. A core in the first layer will allow strain relief in the InAs from the second layer onwards. The cohesive pair energy for biaxially strained InAs is -7.63 eV, but for InAs strained along [001] only it is -7.77 eV. Thus there is an energy difference of 0.14 eV per InAs pair between the two states of strain. Given that there are fourteen pairs per layer we expect a total difference of -1.96 eV between a layer containing biaxially strained InAs and one containing InAs under uniaxial strain. This

estimate neglects the energy of the dislocation cores themselves, but serves to illustrate why we might expect the dislocation to sit in the first layer rather than above. We approach this question by calculating the energy of dislocations with In-centred cores in the first and second layers.

Once we have established the energetically preferred layer we must also discover the preferred mirror-plane for the core. The core may be centred either over an In atom or an As atom, whichever corresponds to the lowest energy configuration. After calculating the preferred layer we answer this question by calculating the energies of cores over both elements. Having thus established the lowest energy core we may also calculate the critical epilayer thickness at which the dislocation network becomes lower in energy than the equivalent amount of coherent epilayers.

6.5.7 Indium Core in First Layer

The dislocation at the first layer is formed by removing a pair of atoms from each InAs epilayer from the coherent structure. The remaining InAs can then relax the compressive strain along $[1\bar{1}0]$ until their lateral separation is $\frac{15}{14}$ that of the GaAs below (since there are now fourteen InAs pairs lying above the fifteen GaAs pairs). As the strain is relaxed the energy of the system is lowered (see table 6.7 showing the decrease in energy between InAs under biaxial strain and uniaxial strain). Previously we showed that a biaxial strain of 7% increased the cohesive pair energy of bulk InAs by 0.2 eV. Now the InAs has relaxed along $[1\bar{1}0]$ direction, but remains compressed along [001]. The pair energy for uniaxially strained InAs is -7.77 eV whereas in the relaxed bulk it is -7.82 eV. Thus there remains a residual strain energy

of 0.05 eV even after formation of the dislocation network.

The cell containing the dislocation at the heterointerface is displayed in figure 6.14. The dipping of the (110) surface above the dislocation core is not easily seen from this perspective, but one can see the depression of the atoms above and below the dislocation core. Note that there is also a small dip in the (110) surface below the core too, as we have not held the bottom of the slab fixed. There are five InAs epilayers on top of seven GaAs substrate layers, five being the maximum number of epilayers we model as experiments showing the presence of a relatively complete dislocation network at this coverage. The atoms at the heterointerface are strained with respect to their neighbours beneath, they are displaced in the direction of the core, filling the void left by the missing pair. The displacement is greatest for those atoms nearest to the core and falls to zero for atoms which are halfway between cores.

The core reconstruction results in the Indium atom being bonded to five arsenic atoms, instead of four, see figure 6.15. The Gallium atom beneath moves down into the substrate, and retains bonds to only three arsenic atoms. The Gallium is also pushed backwards along the [001] direction during the reconstruction, by about 0.3 Å from its original position. We can most clearly see these details from a side on view of the repeating supercell core, showing the Gallium immediately below the core, as in figure 6.16. The five-coordinated In and the three coordinated Ga beneath are in agreement with the model of Oyama et al [53] who also use density functional theory to calculate the relaxed structure of this core position.

Although the In immediately above the dislocation is depressed it is the Gallium atom beneath which is displaced most relative to its neighbours.

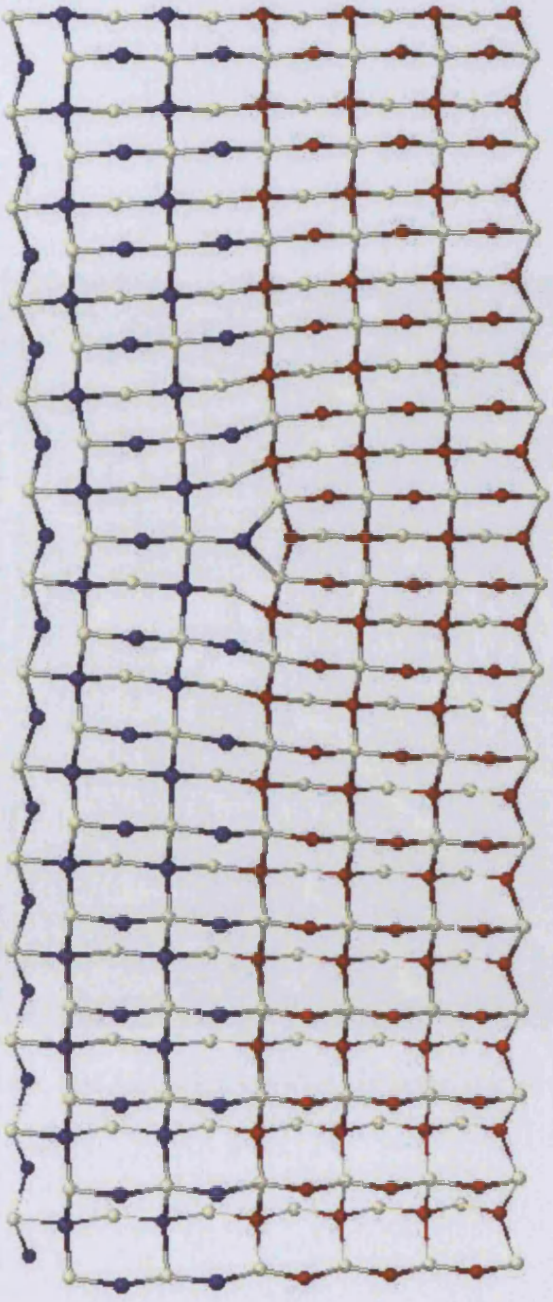

Figure 6.14: InAs/GaAs Cell containing dislocation at the first layer (blue- In, red - Ga, white - As).

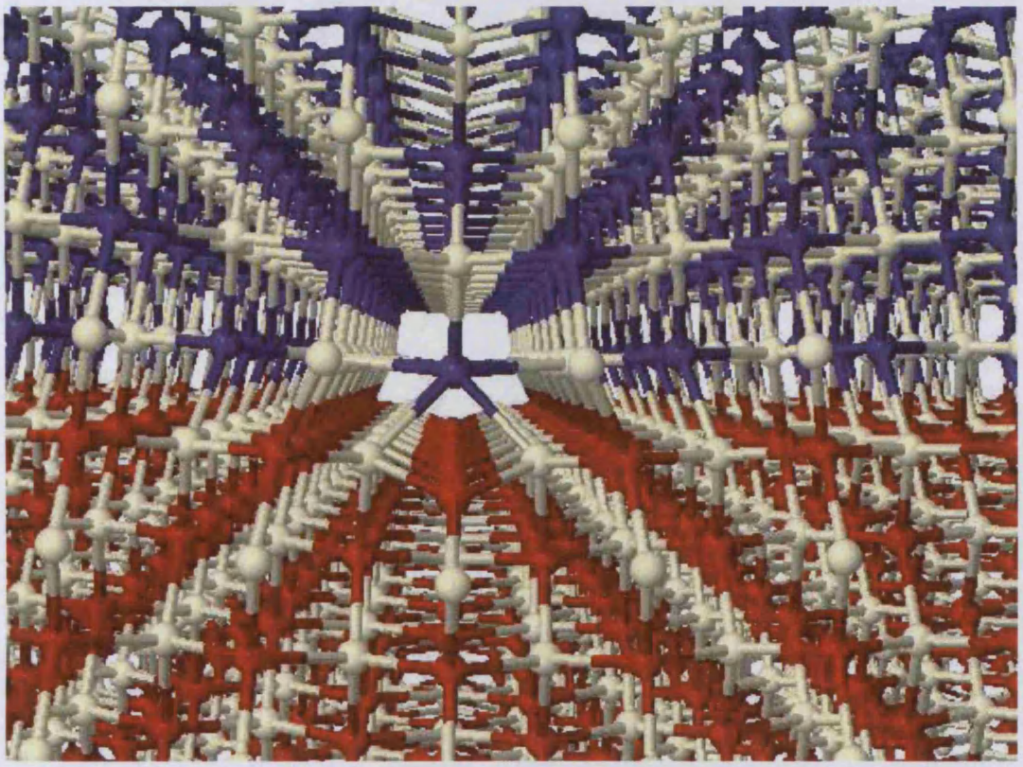

Figure 6.15: Picture of dislocation core at 1st layer (blue- In, red - Ga, white - As).

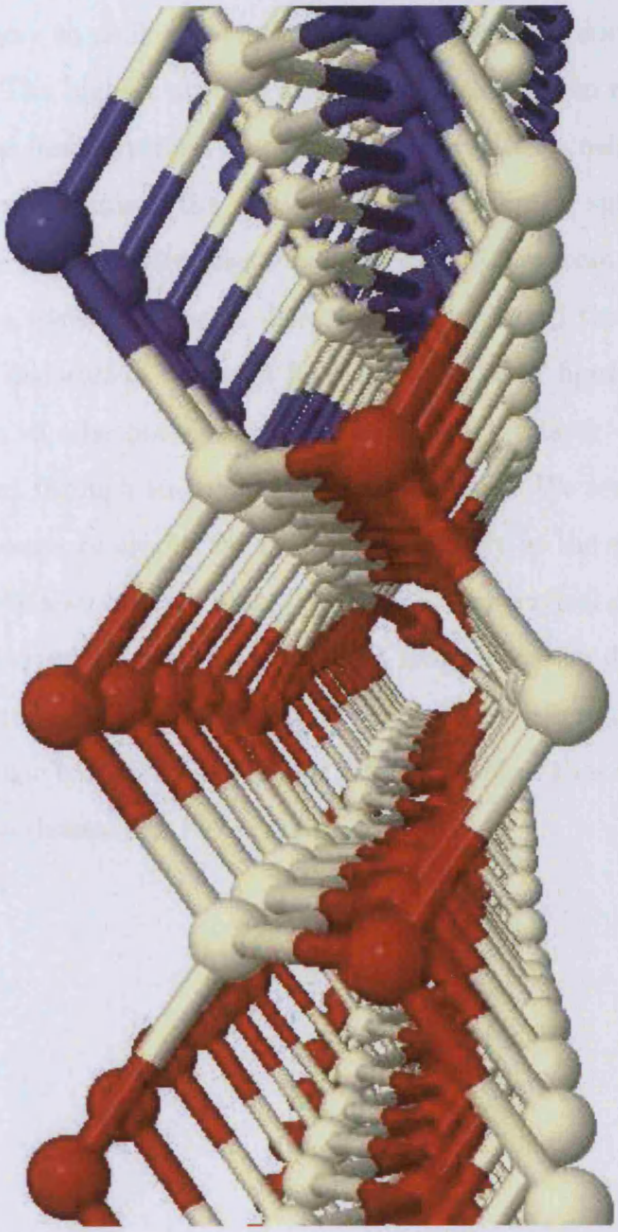

Figure 6.16: Gallium pushed out from its original position.

We measure the relative vertical displacement of an atom in a given row by subtracting its z coordinate from that of the highest atom of the same species in its row (this should be as accurate as subtracting its original z coordinate, in fact even more so as it automatically compensates for any drifting of the surface slab). The highest atom is at the midpoint between repeating cores, and is also the least strained along $[1\bar{1}0]$ relative to its neighbours below. For example, we calculate the size of the surface dip by subtracting the z position of the As atom directly above the dislocation from that of the As furthest from a dislocation core. Below we have plotted the vertical dip of atoms on the dislocation line as a function of layer in figure 6.17. The x axis goes from -6 (the bottom GaAs layer) to 0 (the layer of GaAs at the heterointerface) through to 5, the surface InAs layer. We see the maximum vertical displacements are for the atoms immediately at the dislocation core, as expected. We also see that the magnitude of the vertical dip decreases as we move further up or down the supercell away from the dislocation core. The vertical strain seems to follow a $\frac{1}{d}$ relationship where d is the vertical distance from the core. This behaviour reflects the fact that strain fields fall off with inverse distance in bulk crystal.

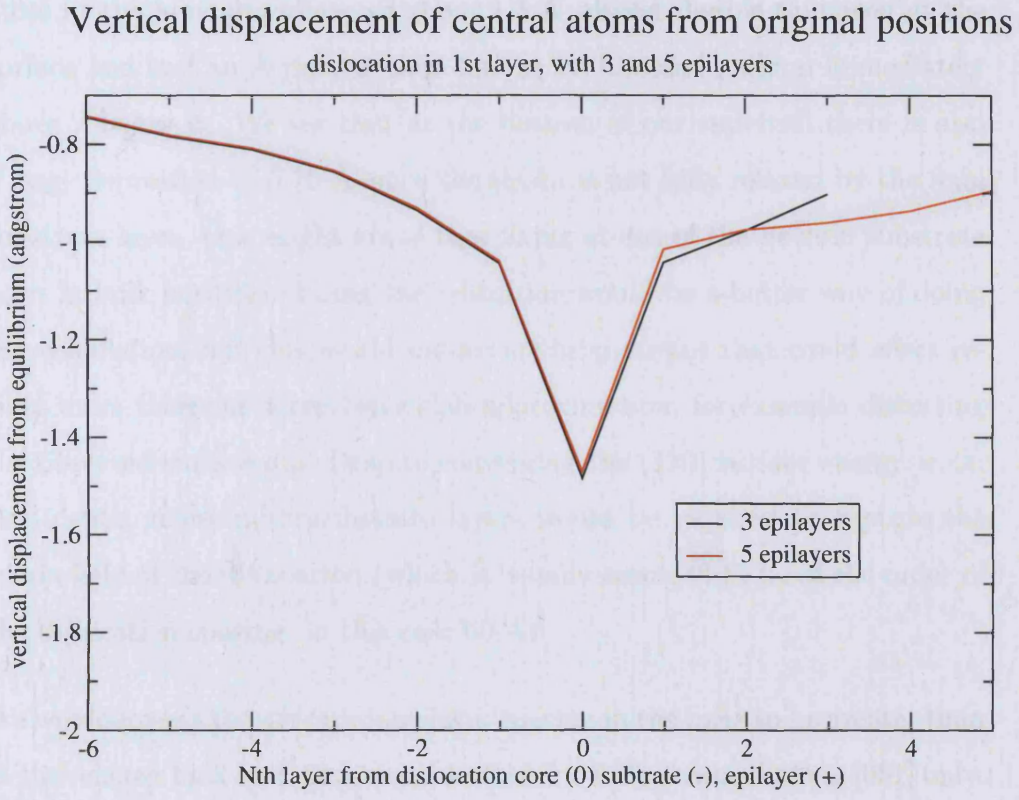

Figure 6.17: Magnitude of the vertical displacement of atoms centred along the dislocation line, red - 5 epilayers, black - three epilayers.

We see that the magnitude of the surface dip is the same for both 3 and 5 epilayers, at 0.9 Å. This is higher than the 0.7 Å measured in [16]. Our measurements correspond to Belk et al's experimental results [12] as they measure a constant 0.7 Å dip at 3 and 5 epilayers too. The small disagreement, rather than being geometrical, may also be due to electronic effects on the STM imaging contrast. The largest displacement is that of the gallium atom at the heterointerface, of about 1.5 Å almost double that seen at the surface and half an Å greater than for the indium and gallium immediately above / below it. We see that at the bottom of our supercell there is also a large depression of 0.75 Å since the strain is not fully relaxed by the final substrate layer. One might argue that fixing atoms of the bottom substrate layer in bulk positions during the relaxation would be a better way of doing the calculation, but this would induce artificial strains that could affect results more than our seven layer slab approximation, for example distorting the observed surface dip. Despite converging the (110) surface energy w.r.t. slab depth, many more substrate layers would be required to capture the strain field of the dislocation (which is usually assumed to be of the order of the dislocation spacing, in this case 60 Å).

We would expect the average interlayer spacing in the InAs to be greater than in the relaxed bulk and correspond to that for InAs strained along [001] only. We measure the average interlayer spacing as a function of epilayer and show the plot below, figure 6.18. We calculate the interlayer spacing averaged across the supercell width, as it is non-uniform due to the dislocation. The spacing increases from just under 2.1 Å between the two layers of GaAs beneath the heterointerface to 2.28 Å between the third and fourth InAs layers above the interface. This agrees well with our value of 2.27 Å for the InAs (110) spacing under uniaxial strain (table 6.7). We do not include the

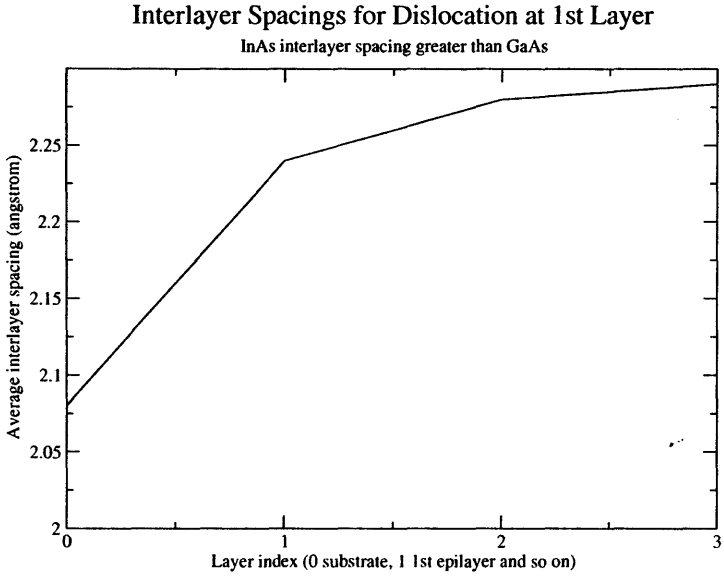

Figure 6.18: InAs interlayer spacing from 1 to 3 epilayers (1st layer core).

spacing between the fourth and fifth epilayers as it is affected by surface relaxation, and therefore irrelevant for comparison with the bulk. There is a 0.16 Å increase in the spacing as we move from the GaAs substrate into the first layer of InAs, and the increments thereafter are much smaller being of the order of hundredths of angstroms rather than tenths.

We also examine the dislocation strain field along $[1\bar{1}0]$, observing the maximum strains at the dislocation core. In this case we measure the lateral strain by measuring the displacement along $[1\bar{1}0]$ between atoms in neighbouring layers. In pseudomorphic growth these atoms lie directly on top of each other, and the lateral strain is zero. However the edge dislocations cause the epilayer atoms to shift from the pseudomorphic positions, resulting in strain relative to the underlying substrate. This is clear from figure 6.14. We show the relative displacements between atoms in the first InAs epilayer and those beneath in figure 6.19 below.

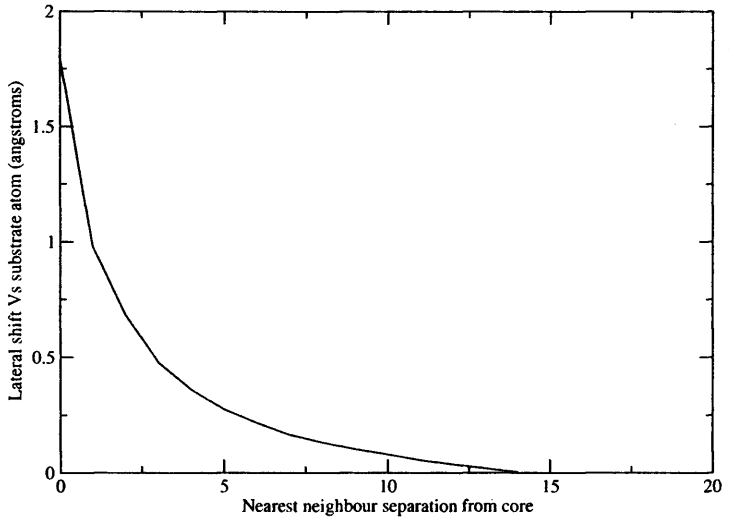

Figure 6.19: $[1\bar{1}0]$ shift of interfacial InAs relative to the GaAs substrate (3 epilayers).

Again we see that the strain field falls to zero for atoms which are midway between repeating dislocations. This is actually imposed by symmetry, since the atoms midway between two dislocations should not be biased toward either of them even with a non-equilibrium dislocation spacing. Indeed, if this were not the case it would reflect a lack of structural relaxation. We also see the same $\epsilon \approx 1/r$ behaviour of the strain field as for the magnitude of the vertical strain with distance from the dislocation core. Again, this is in good agreement with classical elasticity theory.

6.5.8 Geometry of Core at Second Layer

The dislocation core at the second layer above the interface is geometrically similar to the core at the interface. The hallmark five-coordinated In atom with a three coordinated (In) atom below is again observed, as in figure 6.20, which shows the core with five epilayers in total. The InAs bonds at the core

are also longer than the bonds further away, increasing from 2.67 Å up to 2.87 Å. The In immediately below the core moves downwards in a similar manner to the Ga atom in the core at the interface, and dipping is visible at both the top and bottom surfaces.

We measure the dip of the As atom above the second layer core as being 0.9 Å which is the same value we obtained for the first layer core.

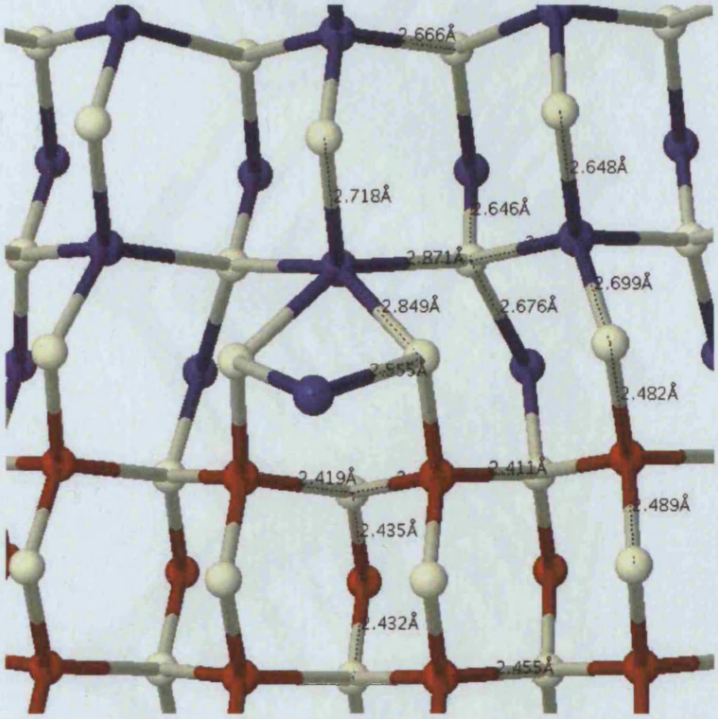

Figure 6.20: Geometry of core at 2nd layer, blue - In, red - Ga, white - As.

6.5.9 Energetically Preferred Layer

We cannot directly compare the energies of the two supercells containing the dislocations in different layers because they contain different numbers of InAs pairs. When forming a dislocation in a given epilayer we must remove

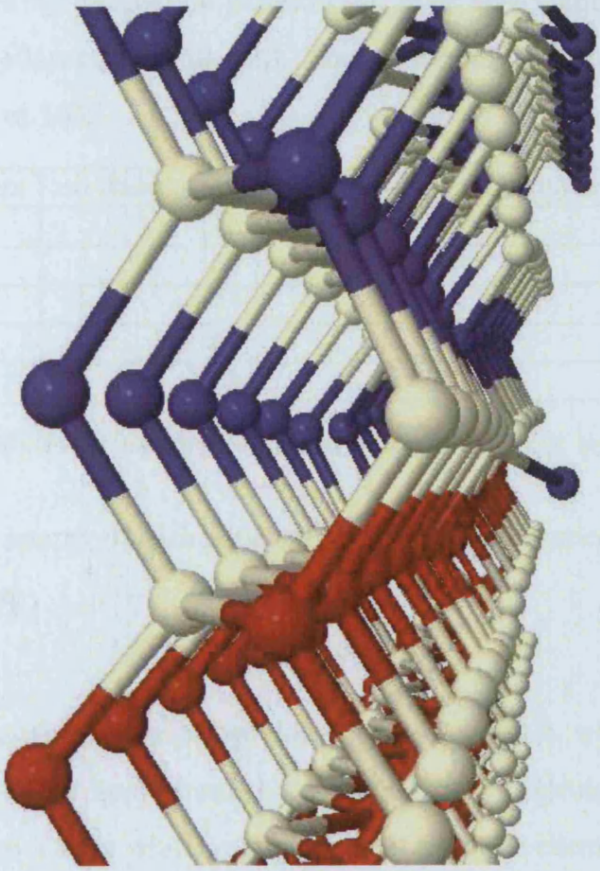

Figure 6.21: Side on view of core at 2nd layer, blue - In, red - Ga, white - As.

an InAs pair from that epilayer and also from all the layers above. The cell with the dislocation at the second epilayer thus contains one more InAs pair than the cell with the dislocation at the heterointerface, and we must compensate for the different numbers of atoms present before comparing the energies. Below we tabulate the number of InAs pairs present in the supercells, indexed by dislocation position (1l, 2l or none), and by the total number of InAs epilayers, in table 6.18. The number of GaAs pairs in all the cells is the same, at 105.

n epilayers	no dislocation	1st layer core	2nd layer core
1	15	14	15
2	30	28	29
3	45	42	43
4	60	56	57
5	75	70	71

Table 6.18: Different numbers of InAs pairs in different supercells.

We can write the energy of each supercell in terms of a chemical potential μ for each InAs pair,

$$E_{cell} = E_{total} - \mu N_{InAs} \tag{6.13}$$

If we then subtract the energy of two different cells, a and b, when expressed in this way, we see that apart from the difference in the calculated supercell energy there is also a term which is a multiple of the InAs chemical potential μ_{InAs}.

$$E_{cell1} - E_{cell2} = E_{total1} + \mu_1 N_1^{InAs} - (E_{total2} + \mu_2 N_2^{InAs}). \tag{6.14}$$

To determine the relative energies of supercells containing different InAs amounts we must decide on the chemical potential μ. The chemical potential of an atomic species must be equal throughout a system which is in

n epilayers	no core	1st layer core	2nd layer core
1	na	na	na
2	-1090.8009	-1074.2343	na
3	-1205.1566	-1182.5994	-1189.6630
4	-1319.9400	-1291.6950	-1298.4340
5	-1434.0182	-1400.4234	-1407.0385

Table 6.19: Table of dislocation cell total energies (eV).

thermodynamic equiibrium, i.e. the relaxed cells. To correct for the different InAs amounts, we need to calculate the energetic cost of adding/ subtracting a single InAs pair from each supercell to obtain the chemical potential of InAs in that particular system. For instance, we wish to compare the energy of a cell with a first layer dislocation and five epilayers against that of one with a second layer dislocation (same epilayers).

We also compare cells containing dislocations against cells without any dislocations. Again, the number of InAs pairs will differ. Below we make both a table and a plot showing the energy of different supercells as a function of the number of epilayers.

Now we must compare the energies of the first and second layer dislocations as a function of the total epilayer number to see which is most stable. In order to compensate for the energy of an InAs pair we use the value -7.77 eV which we calculated in section 6.3.10 for InAs under uniaxial strain, as it is in a similar state of strain to the InAs above the dislocation cores. We have seen that the (110) interlayer spacing agrees well between the uniaxially strained bulk and the InAs above the dislocations. In table 6.20 we compare the energies and find that the core at the first layer is consistently lower in energy than the second layer. The difference is of the order of an eV, increasing from 0.7 eV at three epilayers to 1.15 eV at five. We would expect the difference to

Epilayer	1st layer E	1st layer E + μ	2nd layer E	Δ(1-2)
3	-1182.60	-1190.37	-1189.66	0.71
4	-1291.70	-1299.47	-1298.43	1.04
5	-1400.42	-1408.19	-1407.04	1.15

Table 6.20: Comparing energy (eV) of 1st layer core with 2nd layer core (correction term is -7.77 eV).

converge once a sufficient number of epilayers have been deposited. Thus we have established that the core at the first layer is energetically more stable, in agreement with the conclusions of Oyama et al [53], though they present no explicit calculation of this result.

6.5.10 Dislocation Symmetry Plane

Now we have shown that the first layer is the preferred layer for the dislocation core we must establish the preferred symmetry plane of the core which may lie either on In or As. We construct a simulation cell with a first layer core over As and three epilayers in total and then perform a geometry optimization in order to obtain an energy. However during the course of the optimization rebonding of the dislocation core led to its repositioning over the In atom instead of As. The symmetry may have been broken due to small inaccuracies in the initial atomic positions for example, though these were very small (less than thousandths of an Å). This indicates that the core sits preferentially on In rather than As. Figures 6.22 to 6.25 show the rebonding during the optimization, with the As centred misfit dislocation core eventually rebonding around the neighbouring In atom.

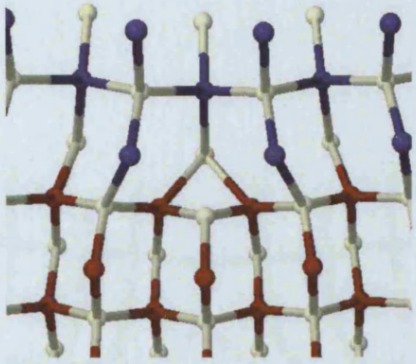

Figure 6.22: Core over As (1). (blue- In, red - Ga, white - As)

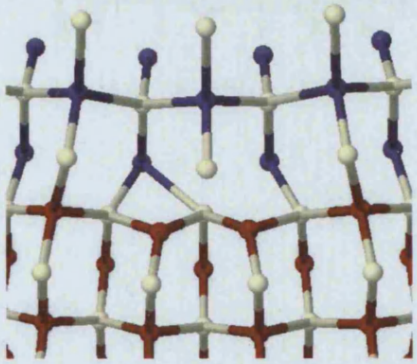

Figure 6.23: Core over As (2). (blue- In, red - Ga, white - As)

6.5.11 Critical Epilayer Thickness

We have shown that the lowest energy edge dislocation network lies in the first layer and is centred on In. Now we can compare the energy of the dislocation network against the energy of the equivalent amount of coherently strained InAs in order to discover the critical epilayer thickness where plastic relaxation of the strained InAs becomes favourable as opposed to continued coherently strained growth. Again the number of InAs pairs in the simulation cells will differ between the coherent InAs and that containing dislocations

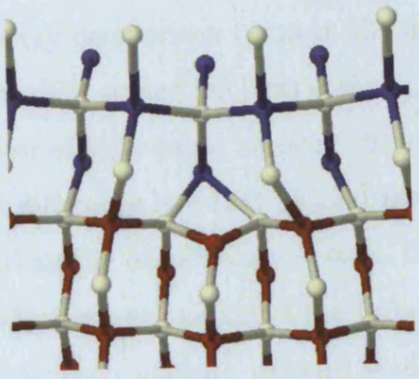

Figure 6.24: Core over As (3). (blue- In, red - Ga, white - As)

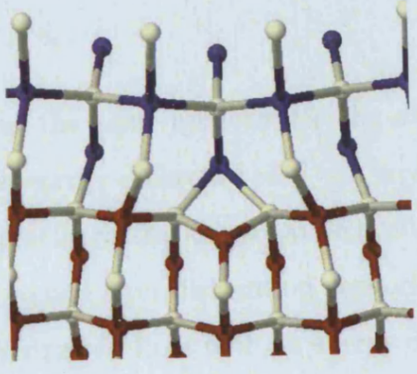

Figure 6.25: Core over As (4). (blue- In, red - Ga, white - As)

(see table 6.18 for the relative numbers). Thus we will use the same compensation value of -7.77 eV per uniaxially strained InAs pair in order to make the comparison.

Table 6.21 shows the energy comparison between the first layer dislocation and the coherent epilayers. The correction term $\Delta E_{1stlayer}$ is simply $m*-7.77$ eV, where m is the number of InAs pairs "missing" from the dislocation cell. At $\theta = 2$ the calculated difference is +1.03 eV and the coherently strained InAs is lower in energy than the dislocation network. However by $\theta = 3$ the difference is -0.75 eV and plastic relaxation of the InAs becomes favourable. As θ is increased the dislocation network remains stable relative to the coherent InAs, which we would expect as the strain relief along $[1\bar{1}0]$ leads to a lowering of the internal energy of the InAs. A simple linear interpolation between $\theta = 2$ and $\theta = 3$ gives us $\theta_{crit} = 2.6$ ML. This differs from the calculated value of Okajima et.al. [56] who derive $\theta_{crit} = 2.35$ ML. However our calculation relies on a more direct numerical approach whereas Okajima et al apply classical elasticity theory in order to interpolate between the energy differences they calculate between the dislocation network and the coherent epilayers at $\theta = 2$ and $\theta = 4$.

In table 6.22 we perform the same analysis for the dislocation over In in the second epilayer. The energy difference at $\theta = 3$ is very small and is not definitive as to whether plastic relaxation would be favoured or not. However by $\theta = 4$ we see that the second layer dislocation network becomes favourable as opposed to coherently strained InAs with an energy difference of -1.80 eV.

n epilayers	$\Delta E_{1stlayer}$	Correction	$\Delta E_{1stlayer}corrected$
2	16.57	-15.54	1.03
3	22.56	-23.31	-0.75
4	28.24	-31.08	-2.84
5	33.59	-38.85	-5.26

Table 6.21: table of 1st layer dislocation energies compared to coherent system (eV).

n epilayers	$\Delta E_{2ndlayer}$	Correction	$\Delta E_{2ndlayer}corrected$
1	na	na	na
2	na	na	na
3	15.49	-15.54	-0.05
4	21.51	-23.31	-1.80
5	26.98	-31.08	-4.10

Table 6.22: table of 2nd layer dislocation energies compared to coherent system (eV).

6.5.12 Conclusions

We have examined the formation of misfit dislocations during the strained heteroepitaxial growth of InAs on GaAs(110) using DFT. The lowest energy structure of the edge-dislocation network which forms to relieve mismatch strain along $[1\bar{1}0]$ is found to be in the first layer rather than the second with a core centred over In. The In at the core is five-coordinated and the Ga beneath is three-coordinated, in agreement with the ab initio calculation of Oyama et.al. [53]. Having established the lowest energy edge-dislocation we then calculate a critical thickness at which formation of the dislocation network becomes energetically favourable. The dislocation network is 1.05 eV higher in energy than the coherent epilayers at $\theta = 2$ but by $\theta = 3$ the network is 0.75 eV lower in energy. Interpolating between these values then gives us a value $\theta_{crit} = 2.6$ ML. Okajima et.al. [56] obtain a slightly different value of 2.35 ML after interpolating between the relative energies of the

coherent and semi-coherent growth modes at $\theta = 2$ and $\theta = 4$, finding energy differences of 3.15 eV/cell and 0.85 eV/cell at each coverage respectively. Our value of θ_{crit} also agrees well with the observations of Belk et.al. that the dislocation network forms at $\theta = 3$ ML. We measure the magnitude of the surface depression due to the dislocation to be 0.9 Å which is close to Belk et.al's value of 0.7 Å (see [12], though a value of 0.5 Å is quoted in [16]). We observe no change in the magnitude of the depression on going from $\theta = 3$ to $\theta = 5$ also in agreement with the observations of Belk et al.

Chapter 7

Conclusions

In this thesis we have demonstrated the successful implementation of a basis of pseudo atomic orbitals (PAOs) within CONQUEST code and used them to examine the performance of the diagonalisation and O(N) algorithms on bulk Si and Si(001).

We produced and compared strain energy curves for bulk Si using direct diagonalisation of the PAOs, obtaining the total energy with both the non self-consistent Harris Foulkes (HF) functional and the Kohn-Sham (KS) self-consistent functional. We found little difference in the energies obtained, with errors of less than 0.05 eV per atom, showing the Harris-Foulkes functional to be a reasonable approximation in bulk Si.

Turning to comparisons of the O(N) algorithm with direct diagonalisation we found that energies obtained using O(N) converged systematically towards the exact diagonalisation value with respect to increasing range of the L matrix, which is used to form the auxiliary density matrix. An L range of greater

than 20 Bohr was required to obtain results close to exact diagonalisation values. However the bulk modulus, which is related to the second derivative of the strain energy curve, did not appear to converge so well as the total energy or equilibrium lattice constant, perhaps due to its being more sensitive to the precise shape of the strain curve. Calculations performed on the Si(001) surface also demonstrated the monotonic convergence of the maximum force from the O(N) algorithm (non self-consistent) towards the (non self-consistent) diagonalisation value.

The presence of the new basis within CONQUEST has led to work on the strained growth of Ge on Si which is currently being written up for publication. In this thesis we have presented a study of the initial stages of strained growth of InAs on GaAs(110). Many approximations were necessary in our calculations in order to bring down the system size to something that was computationally manageable. Order N DFT makes possible the study of much larger systems that correspond more closely to experimental situations. CONQUEST, with the PAO basis, is being applied to analyse the energetics of Ge "hut clusters" which form during strained growth on Si(001) substrate, we show a hut cluster in figure 7.1. These hut clusters contain many thousands of atoms and as such lie beyond the capabilities of conventional DFT codes. Using CONQUEST's O(N) capabilities Ge/Si(001) unit cells containing up to 23000 atoms [5] have been studied (using the NEC Earth Simulator supercomputer in Japan) in work that breaks new ground in the application of DFT to large, complex systems - T. Miyazaki, M. J. Gillan and D. R. Bowler, "A DFT study of strained Ge (105) surface using localized orbital basis sets and a linear-scaling theory", in preparation.

A public release of CONQUEST is now being targeted for the end of this year,

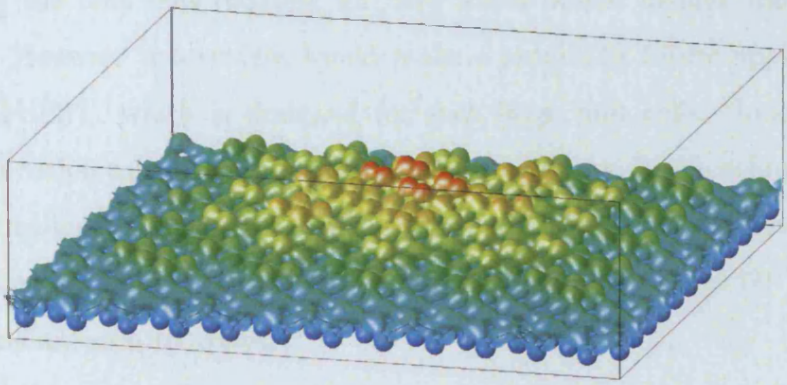

Figure 7.1: Ge on Si(001) hut cluster (4096 atoms).

and the PAO basis will also play an important role in future applications of the code.

Our calculations of misfit dislocations in chapter six have yielded the energetically preferred dislocation position and also the critical InAs epilayer thickness at which their onset is predicted. We find that the dislocation at the first epilayer is lower in energy than that at the second, and that it should have a core on In, since the As core proves unstable. For the first layer core we calculate the critical thickness to be 2.6 ML. This corresponds to the experimental observation that misfit dislocations first begin to appear after 3 ML deposition, as well as previous estimates in the literature. Also in good correspondence with experiment is our calculation that the magnitude of the dip above the dislocation should be unchanged on going from 3 to 5 epilayers, although we estimate a magnitude of 0.9 Å whereas experiments measure 0.7 Å.

The energetics of the 60° degree type dislocations which relieved the resid-

ual strain in the InAs were beyond the scope of our plane-wave DFT approach, as the unit cells required for their study would include thousands of atoms. However this system would make a promising future application for CONQUEST, which is designed for such large unit cells. Indeed the misfit dislocation calculations described here would provide a benchmark for the performance of CONQUEST at low epilayer InAs coverages before it is used in tackling the higher coverages (tens of epilayers) at which 60° degree dislocations are seen to appear.

The promise of CONQUEST is now being realised in its application to systems as complex as Ge hut clusters, and the incorporation of PAOs has been very important in reaching this stage. In future it is hoped that CONQUEST will be used by many groups around the world for modelling systems out of the range of conventional plane-wave DFT, standing alongside other linear scaling codes (SIESTA,OPENMX,ONETEP), with its own unique, efficient and robust linear scaling approach, as well as a choice of basis sets which may be used according to the problem at hand.

Appendix A

Thesis related publications and proceedings

1. "Recent progress with large-scale ab initio calculations: the CONQUEST code", D. R. Bowler, R. Choudhury, M. J. Gillan and T. Miyazaki, phys. stat. sol., 243:989-1000,2006

2. "Large-scale ab initio calculations", T. Miyazaki, R. Choudhury, D. Bowler and M. Gillan, Proc. Int. Conf. Computational Modeling and Simulation of Materials, Acireale, Sicily, 30. May - 4. June 2004

3. "Atomic force algorithms in DFT electronic-structure techniques based on local orbitals", T.Miyazaki, D.R.Bowler, R. Choudhury and M.J.Gillan, Journal of Chemical Physics 121, 6186 (2004)

4. "Misfit dislocation formation during InAs/GaAs(110) heteroepitaxy.", R. Choudhury et al, in preparation.

Bibliography

[1] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136(3B):B864–B871, Nov 1964.

[2] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140(4A):A1133–A1138, Nov 1965.

[3] R.G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules.* Oxford University Press, 1994.

[4] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Mod. Phys.*, 64(4):1045–1097, Oct 1992.

[5] D.R. Bowler, R. Choudhury, M.J. Gillan, and T.M. Miyazaki. Recent progress with large scale ab initio calculations: the CONQUEST code. *Phys. Stat. Sol.*, 243:989–1000, 2006.

[6] Weitao Yang. Direct calculation of electron density in density-functional theory. *Phys. Rev. Lett.*, 66(11):1438–1441, Mar 1991.

[7] Bowler D.R. and Gillan. M.J. Density matrices in O(N) electronic structure calculations: theory and applications. *Comp. Phys. Comm.*, 120:95–108, 1999.

[8] X.P. Li, R. W. Nunes, and D. Vanderbilt. Density-matrix electronic-structure method with linear system-size scaling. *Phys. Rev. B*, 47(16):10891–10894, Apr 1993.

[9] Otto F. Sankey and David J. Niklewski. Ab initio multicenter tight-binding model for molecular-dynamics simulations and other applications in covalent systems. *Phys. Rev. B*, 40(6):3979–3995, Aug 1989.

[10] J.M. Soler, E. Artacho, J.D. Gale, A. Garcia, J. Junquera, P. Ordejon, and D. Sanchez-Portal. The SIESTA method for ab initio order-N materials simulation. *J. Phys. Cond. Matt.*, 14:2745–2779, Nov 2002.

[11] E. Hernández, M. J. Gillan, and C. M. Goringe. Basis functions for linear-scaling first-principles calculations. *Phys. Rev. B*, 55(20):13485–13493, May 1997.

[12] J. Belk. J. Belk PhD Thesis. *PhD Thesis, Imperial College, University of London*, 1997.

[13] R. Martin. *Electronic Structure; basic theory and practical methods.* Cambridge University Press, 2004.

[14] C.M. Goringe, D.R. Bowler, and E. Hernandez. Tight-binding modelling of materials. *Rep. Prog. Phys.*, 60:1447–1512, Apr 1997.

[15] Stefan Goedecker. Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71(4):1085–1123, Jul 1999.

[16] J. G. Belk, J. L. Sudijono, X. M. Zhang, J. H. Neave, T. S. Jones, and B. A. Joyce. Surface Contrast in Two Dimensionally Nucleated Misfit Dislocations in InAs /GaAs(110) Heteroepitaxy. *Phys. Rev. Lett.*, 78(3):475–478, Jan 1997.

[17] R. O. Jones and O. Gunnarsson. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.*, 61(3):689–746, Jul 1989.

[18] A. Szabo and N.S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory.* Dover Publications, 1996.

[19] W. Matthew C. Foulkes and Roger Haydock. Tight-binding models and density-functional theory. *Phys. Rev. B*, 39(17):12520–12536, Jun 1989.

[20] G.B. Arfken and H.J. Weber. *Mathematical Methods for Physicists.* Elsevier Science, 2001.

[21] D. M. Ceperley and B. J. Alder. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.*, 45(7):566–569, Aug 1980.

[22] R. Car and M. Parrinello. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.*, 55(22):2471–2474, Nov 1985.

[23] D. R. Hamann, M. Schlüter, and C. Chiang. Norm-Conserving Pseudopotentials. *Phys. Rev. Lett.*, 43(20):1494–1497, Nov 1979.

[24] Leonard Kleinman and D. M. Bylander. Efficacious Form for Model Pseudopotentials. *Phys. Rev. Lett.*, 48(20):1425–1428, May 1982.

[25] J. C. Slater and G. F. Koster. Simplified LCAO Method for the Periodic Potential Problem. *Phys. Rev.*, 94(6):1498–1524, Jun 1954.

[26] D. J. Chadi. (110) surface atomic structures of covalent and ionic semiconductors. *Phys. Rev. B*, 19(4):2074–2082, Feb 1979.

[27] A.P. Sutton, M.W. Finnis, D.G. Pettifor, and Y. Ohta. The tight-binding bond model. *J. Phys. C.*, 21(35), Jan 1988.

[28] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal. Quantum Monte Carlo simulations of solids. *Rev. Mod. Phys.*, 73(1):33–83, Jan 2001.

[29] E. Hernández and M. J. Gillan. Self-consistent first-principles technique with linear scaling. *Phys. Rev. B*, 51(15):10157–10160, Apr 1995.

[30] C.K. Skylaris, P.D. Haynes, A. Mostofi, and M.C. Payne. Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. *J. Chem. Phys.*, 122:084119, Feb 2005.

[31] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B. Flannery, and M. Metcalf. *Numerical Recipes in Fortran 90.* Cambridge University Press., 1999.

[32] M.J. Gillan. Calculation of the vacancy formation energy in Aluminium. *J. Phys. Cond. Mat.*, 1(4):689–711, Jan 1989.

[33] Walter Kohn. Theory of the Insulating State. *Phys. Rev.*, 133(1A):A171–A181, Jan 1964.

[34] Gregory H. Wannier. Dynamics of Band Electrons in Electric and Magnetic Fields. *Rev. Mod. Phys.*, 34(4):645–655, Oct 1962.

[35] G. Nenciu. Dynamics of band electrons in electric and magnetic fields: rigorous justification of the effective hamiltonians. *Rev. Mod. Phys.*, 63(1):91, Jan 1991.

[36] W. Kohn. Analytic Properties of Bloch Waves and Wannier Functions. *Phys. Rev.*, 115(4):809–821, Aug 1959.

[37] Jacques Des Cloizeaux. Energy Bands and Projection Operators in a Crystal: Analytic and Asymptotic Properties. *Phys. Rev.*, 135(3A):A685–A697, Aug 1964.

[38] N. March, W. Young, and S. Sampanthar. *The Many-Body Problem in Quantum Mechanics.* CUPS, 1967.

[39] S. Goedecker. Decay properties of the finite-temperature density matrix in metals. *Phys. Rev. B*, 58(7):3501–3502, Aug 1998.

[40] Sohrab Ismail-Beigi and T. A. Arias. Locality of the Density Matrix in Metals, Semiconductors, and Insulators. *Phys. Rev. Lett.*, 82(10):2127–2130, Mar 1999.

[41] Giulia Galli and Michele Parrinello. Large scale electronic structure calculations. *Phys. Rev. Lett.*, 69(24):3547–3550, Dec 1992.

[42] E. Hernández, M. J. Gillan, and C. M. Goringe. Linear-scaling density-functional-theory technique: The density-matrix approach. *Phys. Rev. B*, 53(11):7147–7157, Mar 1996.

[43] R. McWeeny. Some Recent Advances in Density Matrix Theory. *Rev. Mod. Phys.*, 32(2):335–369, Apr 1960.

[44] C. A. White, P. Maslen, M.S. Lee, and M. Head-Gordon. The tensor properties of energy gradients within a non-orthogonal basis. *Chem. Phys. Lett.*, 276(1-2):133–138, Sept. 1997.

[45] T.M. Miyazaki, D.R. Bowler, R. Choudhury, and Gillan M.J. Atomic force algorithms in DFT electronic-structure techniques based on local orbitals. *J. Chem. Phys*, 121(13):6186, Oct 2004.

[46] Adam H. R. Palser and David E. Manolopoulos. Canonical purification of the density matrix in electronic-structure theory. *Phys. Rev. B*, 58(19):12704–12711, Nov 1998.

[47] Eduardo Anglada, José M. Soler, Javier Junquera, and Emilio Artacho. Systematic generation of finite-range atomic basis sets for linear-scaling calculations. *Phys. Rev. B*, 66(20):205101, Nov 2002.

[48] E.P. Wigner. *Group Theory and Applications to Quantum Mechanics.* Academic Press., 1997.

[49] D.A. Varshalovich, A.N. Moskalev, and V.K. Khersonskii. *Quantum Theory of Angular Momentum.* World Scientific Publishing Co., 1998.

[50] N. Troullier and José Luriaas Martins. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B*, 43(3):1993–2006, Jan 1991.

[51] G.P. Kerker. Non-singular atomic pseudopotentials for solid state applications. *J. Phys. C.*, 13:L189–L194, Mar 1980.

[52] C. Kittel. *Introduction to Solid State Physics.* Wiley, New York, 1986.

[53] N. Oyama, E. Ohta, K Takeda, K. Shiraishi, and H. Yamaguchi. First Principle Calculations of Misfit Dislocations in InAs/GaAs(110) Heteroepitaxy. *Surf. Sci.*, 433-435:900–903, Aug 1999.

[54] N. Oyama, E. Ohta, T. Kyozaburo, K. Shirasihi, and H. Yamaguchi. First principles calculations on atomic and electronic structures of misfit

dislocations in InAs/GaAs(110) and GaAs/InAs(110) heteroepitaxies. *J. Cryst. Growth.*, 201-202:256–259, May 1999.

[55] L.A. Zepeda-Ruiz, D. Maroudas, and W.H. Weinberg. Theoretical study of the energetics, strain fields, and semicoherent interface structures in layer-by-layer semiconductor heteroepitaxy. *J. Appl. Phys.*, 85:3677–3695, April 1999.

[56] K. Okajima, K. Takeda, N. Oyama, E. Ohta, K. Shiraishi, and Ohno T. Phenomenological Theory of Semiconductor Epitaxial Growth with Misfit-Dislocations. *Jpn. J. Appl. Phys.*, 39:L917–L920, Sep 2000.

[57] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54(16):11169–11186, Oct 1996.

[58] José Luiz A. Alves, Jörk Hebenstreit, and Matthias Scheffler. Calculated atomic structures and electronic properties of GaP, InP, GaAs, and InAs (110) surfaces. *Phys. Rev. B*, 44(12):6188–6198, Sep 1991.

[59] M. Fuchs, M. Bockstedte, E. Pehlke, and M. Scheffler. Pseudopotential study of binding properties of solids within generalized gradient approximations: The role of core-valence exchange correlation. *Phys. Rev. B*, 57(4):2134–2145, Jan 1998.

[60] Yu-Min Juan, Efthimios Kaxiras, and Roy G. Gordon. Use of the generalized gradient approximation in pseudopotential calculations of solids. *Phys. Rev. B*, 51(15):9521–9525, Apr 1995.

[61] Guo-Xin Qian, Richard M. Martin, and D. J. Chadi. First-principles calculations of atomic and electronic structure of the GaAs(110) surface. *Phys. Rev. B*, 37(3):1303–1307, Jan 1988.

[62] R. J. Meyer, C. B. Duke, A. Paton, A. Kahn, E. So, J. L. Yeh, and P. Mark. Dynamical calculation of low-energy electron diffraction intensities from GaAs(110): Influence of boundary conditions, exchange potential, lattice vibrations, and multilayer reconstructions. *Phys. Rev. B*, 19(10):5194–5205, May 1979.

[63] N. Moll, A. Kley, E. Pehlke, and M. Scheffler. GaAs equilibrium crystal shape from first principles. *Phys. Rev. B*, 54(12):8844–8855, Sep 1996.

[64] C. Messmer and J.C. Bilello. The surface energy of Si, GaAs, and GaP. *J. Appl. Phys.*, 52(7):4623–4629, July 1981.

[65] C. Mailhiot and C. B. Duke. Calculation of the atomic geometries of the (110) surfaces of IIIV compound semiconductors. *Surf. Sci.*, 149:366–380, June 1984.